nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



Network analysis of journeys reveals optimum size for New York taxi fleet PAGE 534

DRIVING FORCE

CLINICAL ONCOLOGY

MICROBIOME MATTERS

Harnessing gut bacteria to aid cancer therapies

PAGE 482

CLIMATE CHANGE

ECONOMICS OF EMISSION CUTS

UN goal of 1.5 °C warming offers potential savings

PAGES 498 & 549

EVOLUTION

HEAD START FOR HUMANS

How ecological factors increased human brain size

PAGES 496 & 554

NATURE.COM/NATURE

24 May 2018

Vol. 557, No. 7706

THIS WEEK



EDITORIALS

SALES PITCH Electric cars talked down by vehicle dealers **p.468**

WORLD VIEW EPA transparency rules are anything but robust science **p.469**

KING OF PAIN The long and short of an insect sting **p.471**

Clarity needed on data protection

As a commendable European law on personal data comes into force, the research community must not let excessive caution about data sharing, however understandable, become the default position.

European policymakers have been discussing new rules on data protection for years, and scientists and universities — like everyone else across the continent — are about to see the results. Entering into force on 25 May, a new law known as the General Data Protection Regulation (GDPR), is designed to protect the personal privacy of citizens and will overhaul how personal data are collected, handled, processed and stored. It's a welcome move to safeguard individuals and is the biggest shake-up of data protection in more than 20 years.

However, as this journal has noted before, earlier drafts of the law posed a problem for science and the research community. Of particular concern was the issue of consent — the draft language suggested researchers would be required to seek renewed consent to reuse data collected for a different purpose, which could have introduced delays and made some research impractical. But many in the research community worked relentlessly to warn policymakers of the potential harm. In response, officials put in place rules that exempt research from some of the requirements, provided the proper safeguards are in place. Universities and organizations have introduced plans to make sure they are. The bulk of the work should be done.

The passing of the final GDPR rules is, therefore, a good example of political engagement by researchers and their advocates, and a sensible and informed reaction from policymakers. Those involved, on both sides, deserve great credit. Harmonization of how data can be sourced, stored and used would, in theory, be good for research. It could smooth the difficulties that scientists face when they try to pool analysis of genomic data and tissue samples across national borders. Such sharing could help scientists to organize powerful trials with large numbers of participants.

But although there is some cause for celebration, there are still outstanding issues. And that means that the same researchers and advocates must remain vigilant.

The problem is that individual European countries have been left to decide some issues for themselves — for example, how scientific data can be processed. This flexibility is intended to allow countries to fit the rules around existing systems and different cultures, but it might leave nations out of step. Researchers who work under different systems could struggle to share data with each other. That could lead to delays in negotiations between institutions wanting to create collaborative contracts that enable data sharing.

To help prevent this and to offer a unified approach, academics, industry representatives and patients have been meeting over the past year to distil the complex regulation into a user-friendly guide. This planned code of conduct aims to provide a simple 'how-to' guide for scientists, for example, by explaining differences in the way countries such as Germany and the United Kingdom define 'anonymized' data. The resulting Code of Conduct for Health Research, overseen by the biobank network BBMRI-ERIC (see J.-E. Litton *Nature* **541**, 437; 2017), is almost ready for consultation. But meanwhile, medical research remains vulnerable to unintended consequences of the new law.

That's because, until the code of conduct is in place to offer clear guidance about how to comply with the GDPR, day-to-day decisions on how to interpret the law will be left to individual institutions' legal departments. It would be understandable if they chose to err on the side of caution and place restrictions on sharing data for fear of breaking the law.

Even when the code is finalized, it must still be approved by the European Data Protection Board (EDPB), which has not yet said how organizations can submit such codes for evaluation, or how long the process will take.

"It's important to ensure data can be used with integrity to support valuable research."

Some have argued that delays in the code becoming available could be beneficial, because they would allow the research community to thrash out the details of this complicated area of the law. But others worry that if the process drags on too long, medical research will suffer. What starts as a cautious position on how best to share data in line with the law could drift into normal practice.

That would be a missed opportunity and could risk undermining the good work done so far. Officials on the EDPB must not allow that to happen. The code must be approved and put into practice as soon as possible. It's important to protect people's personal data; but it's also important to ensure data can be used with integrity to support valuable research. ■

Climate costs

A strong financial case for urgent action on greenhouse-gas emissions has now been made.

Published in 1991, an academic paper called 'To slow or not to slow: the economics of the greenhouse effect' is seen as the first attempt to model the economics of global climate change (W. D. Nordhaus *Econ. J.* **101**, 920–937; 1991). Written by the economist Bill Nordhaus, the 18-page study assessed the costs of acting on emissions and the estimated costs of not doing so, and concluded that it was better for the economies of the world to try to address the problem than simply to give up and take the consequences.

Economists and analysts around the world have repeated the exercise many times, most prominently with the British government's Stern Review in 2006 (N. Stern *The Economics of Climate Change: The Stern Review*; Cambridge Univ. Press, 2007). Almost all agree with the original conclusion: it will be much cheaper to spend the money on trying to curb emissions than to pay for the impact of the resulting

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

climate change. But how much cheaper? There's the rub.

A study in *Nature* this week offers the latest and most comprehensive attempt to address this question (see page 549). Marshall Burke and his colleagues modelled the impact of historical temperature changes on the gross domestic product (GDP) of 165 countries between 1960 and 2010. They then ran the model to the end of this century to plot what would happen to GDP according to how much average temperature rise was expected. The results show that greater action to curb temperature will bring greater economic benefits — which are, in reality, avoided damages, measured as the impact on GDP.

Specifically, there is a 75% chance that keeping global warming to a 1.5 °C rise — an aspiration of the Paris climate treaty — will leave the world better off than letting it run to a 2 °C rise. The probable savings: a cumulative US$20-trillion increase in world GDP by the end of the century. (Global GDP in 2016 was about $76 trillion.)

It's fair to say that not all of the world's economists and climate-policy wonks will be content to take the conclusions of the study at face value. Details matter, not least — as with all models — the kinds of assumptions made and data used. The Stern Review, for example, was quickly queried by economists who criticized the way in which it borrowed from the insurance industry and placed great importance on the needs of future generations, who are usually discounted in models of economic impact because it is assumed that they will be considerably richer and so better able to deal with problems than are today's generation.

In that spirit of debate, this week we also publish two — conflicting — opinions on this study in a News & Views Forum (see page 498). It provides a glimpse of the debate already raging in the economics community. One point of contention is how fair it is to simply extrapolate from past trends into the future. As finance experts are keen to point out, past performance is not always a reliable guide to future yields, and in this case it could be that the people of the future will find ways to adapt to a changing climate that are not accounted for in the model. Such adaptation — the development and widespread introduction of drought-resistant crops, for instance — would offer a significant saving, because food prices would not increase so dramatically if harvests are protected.

"Our burning of fossil fuel is writing cheques that our economy can't afford to cash."

Another feature of the new model is that it assumes that climate change, and the extreme weather it is expected to bring, will have a compound impact on the rise of a nation's GDP. Thus, a devastating storm or washed-out summer would affect not just that year's economic performance, but also its performance in subsequent years. Previous studies have taken a more optimistic view that any damage could quickly be compensated for.

There is a more fundamental issue, too. Just how reliable is GDP as a metric? Famously, it assumes that the market price of goods and services fully reflects the costs of their production and use. And the economics of climate change don't always do that: the price of fossil fuels, for one, doesn't take account of the costs associated with future warming.

Like all models, these economic projections will be argued over, worked on and ultimately improved. Scientists can gather the data to help that process, for example by expanding studies of the regional effects of climate change to poorer nations that are already bearing the brunt of the physical and economic impacts. Meanwhile, the arguments for acting on greenhouse-gas emissions, already many and varied, just got a little stronger. Our burning of fossil fuel is writing cheques that our economy can't afford to cash. ■

Road to nowhere

Electric cars are gaining ground fast but face fossil-fuel favouritism in the showroom.

Who killed the electric car? According to the 2006 documentary of that name it was the automobile companies, and especially General Motors (GM), which produced, and then recalled and crushed, thousands of its pioneering EV1 model in the late 1990s. Arguments still rage about the company's true motives (GM insists it was down to high costs), but two decades on from the EV1 with its niche appeal, it's clear that reports of the death of electric vehicles have been greatly exaggerated. Sales in some places are booming. Figures from the Centre of Automotive Management in Gladbach, Germany, show that nearly half of the new vehicles registered in Norway during the first three months of this year were electric. During the same period, China sold more than 142,000 electric vehicles — still just 2% of the total numbers sold, but a large increase on last year.

What drives these sales? According to a study published this week in *Nature Energy* (G. Zarazua de Rubens, L. Noel & B. K. Sovacool *Nature Energy* https://doi.org/10.1038/s41560-018-0152-x; 2018), it's not the sales staff who work at car dealerships — at least not in most nations in Scandinavia. "Do not buy this, it will ruin you," one prospective buyer was told when they asked about an electric car on sale. "Another would-be customer was gently steered away from an electric model because, the sales person wrongly insisted, it would take two days to drive 350 kilometres — roughly the distance between New York City and Washington DC.

We know this because, in these cases, the customers had no intention of buying a car — electric or otherwise. They were undercover university researchers, indulging in a little 'mystery' shopping to test industry attitudes and the barriers that remain to the widespread adoption of new technologies. In this case, the attitude of the sales staff — largely driven by them not knowing as much about the electric models — was hugely influential. The study analysis suggests that it is the most important predictor of the likelihood that a customer will leave having bought an electric car — which the researchers calculated was a dismal 0% in many of the cities they visited.

In all, the researchers underwent 126 shopping experiences in 82 car dealerships across Denmark, Finland, Iceland, Norway and Sweden. (The ethics of mystery-shopping exercises have been questioned — they waste the time and money of the targets — so the researchers did not spend more than about ten minutes talking to the sales staff in each case.) They conclude that dealers were dismissive of electric vehicles and misinformed shoppers about vehicle specifications. In many cases, it took persistent questions from the mystery shoppers to get the electric-car dealers just to admit that yes, they did actually sell electric cars.

Why would car sales staff make it so difficult for customers to buy a car? Because they want them to buy a different kind of car. As the researchers point out, dealers "strongly oriented customers towards petrol and diesel vehicle options" on sale alongside the electric versions. And that behaviour is typical. The researchers argue that the attitude "mirrors industry and government favouritism towards conventional cars".

Why does this matter? Electric cars are an important strategy for sustainable transport and have reached the point where sales to early adopters must start to give way to sales to a larger "early majority" (J. Lynes *Nature Energy* https://doi.org/10.1038/s41560-018-0173-5; 2018). There is a well-known and much-feared chasm between the two stages, and one that policymakers are trying to bridge with incentives such as subsidies, investment in infrastructure and privileged access to road space (such as allowing electric cars into lanes banned to other cars carrying no passengers). The mystery-shopping study highlights a new and important part of the bridge. Attitudes and incentives in dealerships must be changed — even simple steps such as better training and offering higher commission on successful sales of electric vehicles could help. ■

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.


VINCENT VERDI

Transparency rule is a Trojan Horse

The US Environmental Protection Agency is co-opting scientific trappings to sow doubt, warns **Naomi Oreskes**.

Last month, the US Environmental Protection Agency (EPA) proposed a new rule to "ensure that the regulatory science underlying Agency actions is fully transparent, and that underlying scientific information is publicly available in a manner sufficient for independent validation". The alleged justification is a crisis in science over replicability and reproducibility.

At face value, the proposal might seem reasonable. It isn't.

Many EPA watchers believe that the rule targets long-term epidemiological studies that linked air pollution to shorter lives and were used to justify air-quality regulations. In my view, the rule could keep that and other high-quality evidence from being used to shape regulations, even if there are legitimate reasons, such as patient privacy, why some data cannot be made public. It could potentially retroactively exclude an enormous amount of respected evidence. This would make the EPA less able to serve its function "to protect human health and the environment". The window for speaking up is closing fast.

There is a crisis in US science, but it is not the one claimed by advocates for the rule. The crisis is the attempt to discredit scientific findings that threaten powerful corporate interests. The EPA is following a pattern that I and others have documented in regard to tobacco smoke, pollution, climate, and more. One tactic exploits the idea of scientific uncertainty to imply there is no scientific consensus. Another, seen in the latest efforts, insinuates that relevant research might be flawed. To add insult to injury, those using these tactics claim to be defending science.

ROBUST SCIENCE IS BEING CHALLENGED AND NEEDS TO BE DEFENDED.

A previous attempt to restrict the use of vetted science was the 2001 Data Quality Act, which resulted in guidelines for how information could be used and disseminated. The Competitive Enterprise Institute, a think tank in Washington DC dedicated to limited government, was quick to invoke it to try to prevent the distribution of a major EPA report on climate change.

Those lobbying for data that underlie regulations to be publicly available have not made similar demands for other data, such as the composition of fracking fluids, or the information confidentially supplied by companies to register pesticides with the EPA.

Guests present when the EPA administrator unveiled the rule included US Congressman Lamar Smith (Republican, Texas), who has repeatedly introduced legislation to exclude research justifying the US Clean Air Act; Myron Ebell of the Competitive Enterprise Institute, who has long challenged the scientific consensus on climate change; and lawyer Steve Milloy, who also disputes anthropogenic global warming and has long ties to the tobacco industry, which floated similar proposals in the 1990s to try to thwart regulation of second-hand smoke.

Conspicuously absent were the scientific organizations that are working to improve data and transparency in research. Three prominent journals (*Nature*, the *Proceedings of the National Academy of Sciences*, and *Science*) issued a joint statement condemning the rule, even though new policies at the journals had been used to justify it.

One week before the proposed rule was announced, a group called the National Association of Scholars — which uses an acronym easily confused with that of the prestigious National Academy of Sciences — released a report called 'The Irreproducibility Crisis of Modern Science'. The association (I will not call it the NAS!) describes itself as dedicated to "academic freedom and disinterested scholarship" and has focused mostly on critiquing undergraduate courses. The report dwells frequently on climate science, yet the greatest concerns among scientists over reproducibility relate to biomedicine and psychology. The association's president, anthropologist Peter Wood, has compared climate scientists to circus hucksters. Smith championed the report's launch.

I urge the scientific community to get out ahead of efforts that I believe are intended to exploit discussions about reproducibility and transparency for political ends. For starters, researchers should recognize that the term 'regulatory science' as used in the rule does not carve out some separate category of work commissioned by government agencies. It applies to all science.

The geochemists, hydrologists and forest ecologists who worked out the cause of acid rain in the 1960s and 1970s did not set out to study air pollution. They did not think of themselves as environmentalists, although they became embroiled in a public-policy debate. If regulators had ignored ground-breaking papers, they would not have acted to control acid rain. Something similar can be said for the atmospheric chemists in the 1970s and 1980s who realized that chlorofluorocarbons were depleting the ozone layer. That led eventually to the Montreal Protocol, once decried by industry groups — now hailed as a success.

Even the EPA's own science advisory board has protested that the rule seems to have been designed without input from the scientific community, that public access to data from older studies might not be feasible, and that consideration must be given to confidentiality and privacy and, for complex existing data sets, the cost and effort of making them accessible. The rule also fails to take into account extant government mechanisms for vetting science or ways to conduct independent re-analyses without publicly releasing subjects' personal data.

Robust science is being challenged and needs to be defended. Scrutiny is appropriate. Making it more difficult to apply science to governmental regulations is not.

The comment period on the proposed rule closes soon (30 May). Scientists can submit comments at go.nature.com/2ioiz46. Do. ■

Naomi Oreskes *is a historian of science at Harvard University in Cambridge, Massachusetts, and co-author of the book* Merchants of Doubt.
e-mail: oreskes@fas.harvard.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

SEVEN DAYS *The news in brief*

POLICY

Data dissent

A panel of science advisers to the US Environmental Protection Agency (EPA) has sharply criticized a proposed rule that would limit the agency's use of research based on underlying data that are not publicly available. In a 12 May memorandum, a working group of the EPA Science Advisory Board said that the proposed rule could have broad ramifications and should be submitted to the full advisory board for review. The working group said that the Trump administration failed to explain the benefits of the proposal, did not properly assess its potential regulatory impacts, and failed to answer practical questions about its implementation. See page 469 for more.

Asset uncertainties

The Australian Academy of Science is concerned about the lack of detail in the government's plan to spend Aus$393 million (US$297 million) on research facilities over the next 5 years. On 15 May, the government announced its research infrastructure investment plan. Money will be spent on infrastructure such as next-generation microscopes and imaging equipment, advanced nanoscale fabrication facilities and Australia's nuclear-research capacities. But the academy is concerned that there is limited detail on when money will be available, and that crucial infrastructure upgrades could be several years away.

ANNOUNCEMENT

Following discussions at the World Health Organization last week, it has become clear that the outbreak of Ebola in the Democratic Republic of the Congo requires prompt responses by the research community to help minimize the considerable risks of spread.

Accordingly, *Nature*'s editors and those of other appropriate Nature and Nature Research journals will treat relevant submissions with priority. On publication, *Nature*'s publishers will make the research freely available for a period subject to future review.



PETER POWELL/EPA-EFE/REX/SHUTTERSTOCK

UK eyes EU science scheme

The UK prime minister has declared that she wants her country to be part of the European Union's next major research-funding programme, which will launch after Britain leaves the EU in March 2019. In a speech on 21 May, Theresa May (pictured) said the government would like the option to "fully associate" itself with Horizon Europe, which will run from 2021 to 2027 and will be the European Commission's largest-ever science-funding scheme, with a planned budget of almost €100 billion (US$118 billion). May said that in return for paying to take part, she wanted to influence the shape of the scheme. Thirteen non-EU countries are associate members of the current programme, Horizon 2020, and pay a proportion of their gross domestic product so that their scientists can bid for European grants. But none of these nations has input on the programme's priorities.

RESEARCH

Heart regeneration

Japanese physicians are poised to conduct the world's first clinical study using induced pluripotent stem (iPS) cells to treat heart disease. On 16 May, a health ministry panel approved a plan by Osaka University doctors to use the cells to treat ischemic cardiomyopathy, in which reduced blood flow to the heart compromises the organ's ability to pump blood. Heart-muscle cells created from iPS cells will be grown in 0.1-millimetre-thick sheets that are expected to release growth factors that will regenerate the heart muscle. iPS cells have been used in several trials to treat retinal disease in humans, but at least one of those has been halted after a patient had an adverse reaction to the transplant. The heart trial is expected to start in three people by March 2019.

EVENTS

Volcano explosion

The summit crater of Hawaii's Kilauea volcano exploded on 17 May, sending a plume of ash about 9 kilometres into the air. The explosion happened after two weeks of increased activity, during which lava had spewed through volcanic fissures and triggered earthquakes that reached magnitude 6.9. The volcanic activity has damaged at least 36 structures and prompted more than 1,800 evacuations. Volcanic air pollution has been reported in the southern part of Hawaii's Big Island. The US Geological Survey warned the eruption could become more violent and eject "ballistic" rocks up to 2 metres in diameter. See page 477 for more.

Study on hold

The US National Institutes of Health (NIH) has halted a US$100-million study into the health effects of drinking, after news reports suggested that the alcohol industry might have funded it improperly. On 17 May, NIH director Francis Collins said that he was putting the study on hold

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

while his agency investigated allegations that its officials had courted funding from the alcohol industry in violation of NIH rules. The Moderate Alcohol and Cardiovascular Health trial was being run at Harvard University's Beth Israel Deaconess Medical Center in Boston, Massachusetts, and intended to recruit around 7,800 adults for a 10-year study into whether a single drink per day could confer health benefits.

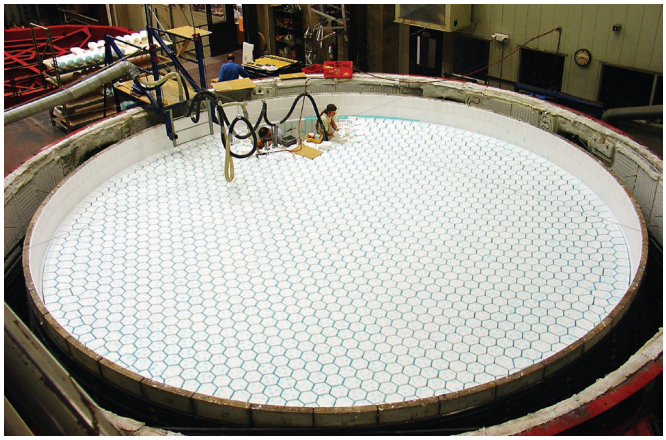
UK visas refused

More than 3,500 scientists and medical and engineering professionals were refused visas to work in the United Kingdom between December 2017 and March 2018. Applicants were refused despite being eligible because the government had, for the fifth month in a row, breached its cap on a class of visa for skilled workers known as Tier 2. Because candidates for PhD-level roles were prioritized for visas, academic scientists were largely unaffected. The government released the figures on 16 May in response to a freedom-of-information request.

SPACE

Observatory pact

Two rival telescope organizations have joined forces to give US astronomers broader access to the next generation of big



UNIV. ARIZONA/SPL

ground-based observatories. The partnership involves the Giant Magellan Telescope Organization (**pictured**, one of its mirrors), which is building a 24.5-metre telescope in Chile, and the Thirty Meter Telescope International Observatory, which aims to construct in either Hawaii or the Canary Islands. They, along with the National Optical Astronomy Observatory, will encourage US astronomers to propose collaborative research programmes for the extremely large telescopes, which the US National Science Foundation could consider funding.

PEOPLE

Energy appointment

US President Donald Trump nominated Christopher Fall to lead the Department of Energy's Office of Science on 18 May. Fall is currently the principal deputy director of the department's Advance Research Projects Agency–Energy. Before that, he was the acting chief scientist at the Office of Naval Research and assistant director for defence programmes at the White House Office of Science and Technology Policy. Fall has a PhD in neuroscience and a master's in business administration. At the Office of Science, he will oversee a research budget of around US$5.4 billion.

Russian minister

The Russian government has divided its science and education ministry in a shake-up after national elections in March, which handed Vladimir Putin his fourth term as president. Prime Minister Dmitry Medvedev appointed the government's new cabinet on 18 May, putting Mikhail Kotyukov in charge of the newly created Ministry of Science and Higher Education. Olga Vasilyeva, a church historian who became head of the now-defunct Ministry of Science and Education in 2016, will head the new Ministry of Education. Kotyukov was previously chief of the Federal Agency for Scientific Organizations, which was established in 2013 and reported directly to Putin. The agency, which will now be dismantled, was controversial because it was given effective control of the Russian Academy of Sciences.

HEALTH

Ebola vaccine

Health officials in the Democratic Republic of the Congo (DRC) began to distribute an experimental vaccine against Ebola on 21 May, as an outbreak of the infection continues to worsen. Forty-six people had been infected with the virus and 26 had died as of 21 May, according to the World Health Organization. An expert committee assembled by the agency says that the health risk from Ebola in the DRC is now "very high". See page 475 for more.

Migraine medicine

The US Food and Drug Administration has approved the first in an emerging class of drugs intended to prevent migraines. On 17 May, the agency announced the approval of Aimovig (erenumab-aooe), an antibody therapy that blocks a protein called calcitonin gene-related peptide receptor. This molecule is involved in pain perception, and other migraine treatments that target it are following close behind. Amgen of Thousand Oaks, California, developed the drug with Novartis of Basel, Switzerland.

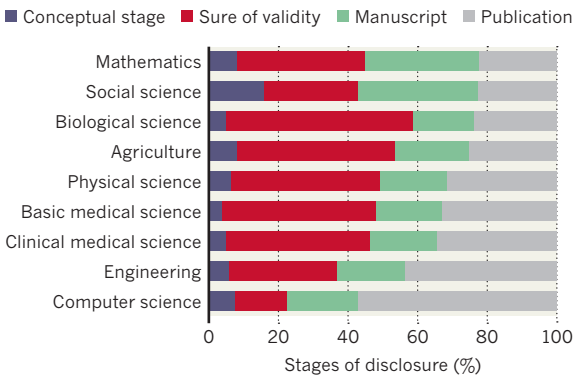
NATURE.COM
For daily news updates see:
www.nature.com/news

TREND WATCH

Two-thirds of scientists share results outside their trusted circle before formally publishing them, a survey of more than 7,000 researchers finds. Social scientists and mathematicians are most likely to disclose findings before publication, and computer scientists least likely. About 40% of scientists disclose a result after they are sure of its validity, and 21% share findings once they've written or submitted a manuscript. The respondents worked in nine fields in the United States, Germany and Switzerland.

SHARING WITH OTHERS

A survey of more than 7,000 researchers finds that about 67% of scientists disclose their results outside their circle of trusted colleagues before formal publication — but when varies widely by field.



SOURCE: J. G. THURSBY, C. HAEUSSLER, M. C. THURSBY & L. JIANG *SCI. ADV.* **4**, EAAR2133 (2018).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

NEWS IN FOCUS

INDONESIA Government plan to clamp down on foreign scientists draws protest **p.476**

HAWAIIAN VOLCANO Scientists scramble to analyse data from Kilauea eruption **p.477**

PUBLISHING Open-access drive escalates as stand-offs spread across Europe **p.479**

CANCER Researchers explore using the microbiome against tumours **p.482**

KENNY KATOMBE/REUTERS



Health Ministry officials carry the first batch of experimental Ebola vaccines in Kinshasa, Democratic Republic of the Congo, on 16 May 2018.

PUBLIC HEALTH

Experimental drugs poised for use in Ebola outbreak

International health groups are in discussions with the Democratic Republic of the Congo.

BY ERIKA CHECK HAYDEN

Aid workers responding to the Ebola virus outbreak in the Democratic Republic of the Congo (DRC) are seeking approval to treat patients with experimental drugs. These include three potential treatments — ZMapp, favipiravir and GS-5734 — that were given to patients during the 2014–16 Ebola epidemic in West Africa.

The drugs are being considered in addition to the use of an experimental vaccine; none of the treatments has been definitively proved to lower the risk of death from Ebola.

The move to test experimental drugs and vaccines early in the outbreak, which was confirmed on 8 May, is part of a push to start research as soon as possible after Ebola cases are detected to save lives. That's a change from the past, when doing research during an outbreak was seen as a distraction.

The switch has been driven by the availability of new vaccines and drugs — and by memories of the 2014–16 epidemic. Officials were so slow to deploy potential vaccines and drugs that the epidemic had waned before clinical trials could start. Now, "there's an acceptance that research during an outbreak is something we need to do", says Daniel Bausch, director of the UK Public Health Rapid Support Team in London.

The DRC allowed the use of an experimental Ebola vaccine during the country's last outbreak, in May 2017, although the outbreak ended before the vaccine was shipped. Earlier this month, the government approved the first shipments of the vaccine and the country now has more than 7,500 doses on hand. The vaccine was given to health-care workers beginning on 21 May, and could be administered this week to patients and their contacts.

Public-health experts hope that the experimental vaccine, called rVSV-ZEBOV, will help to control the outbreak. Forty-six people have been infected and 26 have died, the World Health Organization (WHO) said on 21 May. The virus has spread over a wide area and infected at least one person in a major city, Mbandaka — home to 1.2 million people.

The rVSV-ZEBOV vaccine, manufactured by Merck, was shown to protect against Ebola in a trial run during the West African epidemic. None of the 5,837 volunteers who took the vaccine in that trial became infected.

Officials in the DRC have quashed eight ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

previous outbreaks through conventional public-health measures, such as tracking down people with Ebola and their contacts to understand the disease's path. But they worry about how far the virus has already travelled this time and about the possibility that it could spread even farther — as it did in the West African epidemic, which took root in 3 countries and claimed more than 11,000 lives. "We think the outbreak could become complicated, as it did in West Africa, so we must do everything to stop it," says Jean-Jacques Muyembe-Tamfum, director-general of the National Institute for Biomedical Research in Kinshasa.

Whether that will include deploying experimental drugs is now under discussion. The WHO is consulting experts to consider the evidence for such treatments, and the medical humanitarian organization Médecins Sans Frontières (MSF) is talking to DRC officials about using them, says Annick Antierens, who coordinates Ebola clinical trials for MSF.

Although the rVSV-ZEBOV vaccine could help prevent new infections, Antierens says, experimental drugs might still be needed because officials lack a good understanding of where Ebola first emerged during this outbreak or how it is spreading. So there are likely to be very many people who are already infected.

Using experimental vaccines and drugs in an outbreak raises logistical and ethical complexities, such as delivering them to remote settings by aeroplane or motorbike and designing humane and rigorous clinical trials. The 2014–16 Ebola outbreak saw controversy over whether potential drugs and vaccines should be tested in trials that randomly assign patients to receive either the experimental treatment or standard care. MSF and officials at the WHO argued that withholding the medicines from desperate patients would be unethical.

Some of the treatments MSF is now considering were given to varying numbers of people in the 2014–16 epidemic. ZMapp, an antibody treatment made by Mapp Biopharmaceutical in San Diego, California, was tested in a 72-person trial; 22% of the 36 people who received the drug died, compared with 37% of the 35 who did not receive it. Favipiravir, an antiviral drug from the Japanese company Toyama Chemical, was given to 126 patients in the West African outbreak, and a few dozen in other trials. The antiviral drug GS-5734 was given to three people.

The Congolese Ministry of Health and a national ethics review board would need to approve new drug trials. Observers say that studies must proceed more equitably than they did in the 2014–16 outbreak, when experimental treatments were given first to international doctors and aid workers.

"We were pretty tone deaf," says Lawrence Gostin, director of the WHO Collaborating Center on Public Health Law and Human Rights at Georgetown University in Washington DC. "We need to do that completely differently this time." ■

POLICY

Indonesia plans strict foreign research laws

Regulations could hamper international collaborations.

BY DYNA ROCHMYANINGSIH

Scientists in Indonesia fear that a government plan to introduce stringent rules for foreign researchers will scare off potential collaborators and hamper experiments. The proposals also suggest tough new penalties, including prison sentences, for foreign scientists who break some existing rules, such as the requirement to have a research permit.

Next month, representatives from two science academies will meet with politicians in the hope of convincing them to reconsider the proposals.

"The new regulations will only repel foreign scientists to do research in Indonesia, and this is not good for Indonesia's science," says Berry Juliandi, a member of the Young Academy of Sciences and a biologist at Bogor Agricultural University. The contribution of international scientists is crucial for Indonesian research because foreign science agencies have larger budgets and more sophisticated technology, he says.

Government documents state that the proposed regulations for international science are meant to protect Indonesia's natural resources and to increase local science capacity. The proposals are among several outlined in a draft law submitted to the House of Representatives in August 2017.

If the law is approved by the house, international scientists will have to submit their raw data to the research ministry; involve Indonesian colleagues as equal partners in research projects; and name all Indonesian researchers involved in a project on every peer-reviewed paper that arises from the work.

The draft law also imposes harsh penalties on foreign researchers who break existing regulations. Foreign scientists will still need a government permit to do research, and a special transfer agreement to remove specimens, but breaking these rules would be upgraded to a criminal offence. Researchers could face a prison sentence of up to 2 years, or hefty fines of as much as 2 billion Indonesian rupiah (US$143,000). The current penalty for a researcher who violates a permit can vary from a verbal warning to the permit being revoked. There has been no national policy or penalty for scientists who remove specimens without an agreement.

The draft law would also require that international scientists do research that produces "beneficial output for Indonesia".

Erik Meijaard, a conservation scientist at the University of Queensland in St Lucia, Australia, who studies orangutans in Borneo, says the proposal is "unworkable" for foreigners: "You could do a few years' research, find out that the outcomes do not benefit Indonesia, and then you cannot publish."

Meijaard adds that, overall, the draft law seems vague and is "certain to turn away foreign researchers and stop people from studying in Indonesia if there is an unclear risk of being fined or sent to jail".

The ministry's director-general for research and development reinforcement, Muhammad Dimyati, doesn't think the rules will stop collaboration. "We encourage foreign scientists to publish their research conducted in Indonesia. But they should not write alone and they have to involve Indonesian scientists. This will certainly give benefits to Indonesia's science."

Dimyati says every country has a right to protect its natural resources for the welfare of its people. "The sanctions are intended to remind scientists about their role in society, which is to find innovation that is beneficial for mankind without violating regulations of a country," he says.

"We encourage foreign scientists to publish research conducted in Indonesia. But they should not write alone."

Not all researchers think the penalties are a bad idea. Laksana Handoko, a physicist at the Indonesian Institute of Sciences in Jakarta, supports criminal punishments for researchers who take specimens out of the country without a transfer agreement. It's stealing, he says.

Jason von Meding of the University of Newcastle in Australia, who studies disaster risk-reduction in southeast Asia, says that scientists in developing countries need protection for their work — and that international scientists should not be afraid of the draft law if they've done nothing wrong. But he thinks that less-severe sanctions would be more appropriate than the proposed criminal penalties.

The research ministry's director of intellectual property, Sadjuga, says it could be some time before the proposals become law, because members of the house have to debate the draft, and they are preparing for an election in 2019. ■

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



USGS-HVO

Geological activity around the Kilauea volcano in Hawaii includes several fissures that have opened and begun oozing lava.

GEOLOGY

Hawaii volcano holds clues to other eruptions

Scientists scramble to analyse data from steam explosion at Kilauea.

BY SARA REARDON

After weeks of unleashing earthquakes and lava flows that have forced thousands of people to evacuate their homes, Hawaii's Kilauea volcano finally blew its top last week. Because Kilauea is one of the best-monitored volcanoes in the world, scientists hope that data from the event will help them to better predict when similar volcanoes are about to erupt.

"We'll be working on this set of data for our careers," says Michael Poland, a geophysicist at the US Geological Survey (USGS) Cascades Volcano Observatory in Vancouver, Washington.

The USGS says that the eruption began at 4:15 a.m. local time on 17 May, when the volcano sent a plume of ash and steam more than 9,100 metres into the air.

The many instruments on and around Kilauea were watching. The volcano bristles with equipment that continuously measures signs of geological activity, such as ground movement, lava chemistry and seismic vibrations.

The first hint of an impending eruption came with a series of earthquakes on 3 May. Soon after, fissures opened up in the ground as far as 40 kilometres away from the volcano's rim — oozing lava that forced about 2,000 people to evacuate. The openings also depressurized the network of underground channels beneath Kilauea, including its lava chamber. As a result, the lava level within the volcano's crater quickly dropped by more than 30 metres. It was, Poland says, "like someone pulled the plug in a bathtub".

That caused the walls of the volcano to begin crumbling into the crater, creating a layer of rock atop the surface of the remaining lava. And once the surface of the still-draining lava dropped below the water table, water began to seep into the crater, creating steam and pressure beneath the freshly formed rock cap.

Scientists at the USGS's nearby Hawaiian Volcano Observatory suspected that a steam explosion was imminent: in 1924, the same pattern of oozing fissures around Kilauea had heralded a series of explosions. The researchers were proved right on 17 May, when the pressure that built up in the crater sent debris and ash flying. But Poland says that steam explosions are hard to anticipate.

The hope this time, he adds, is that the extensive instrumentation on Kilauea and data collected from the latest eruption will allow scientists to develop better markers for predicting when a steam explosion is imminent.

For now, scientists in Hawaii and around the world are watching Kilauea and waiting to see what else the volcano will do. "This has so far been playing out how USGS said it would," says Janine Krippner, a volcanologist at Concord University in Athens, West Virginia. "That itself is incredible."

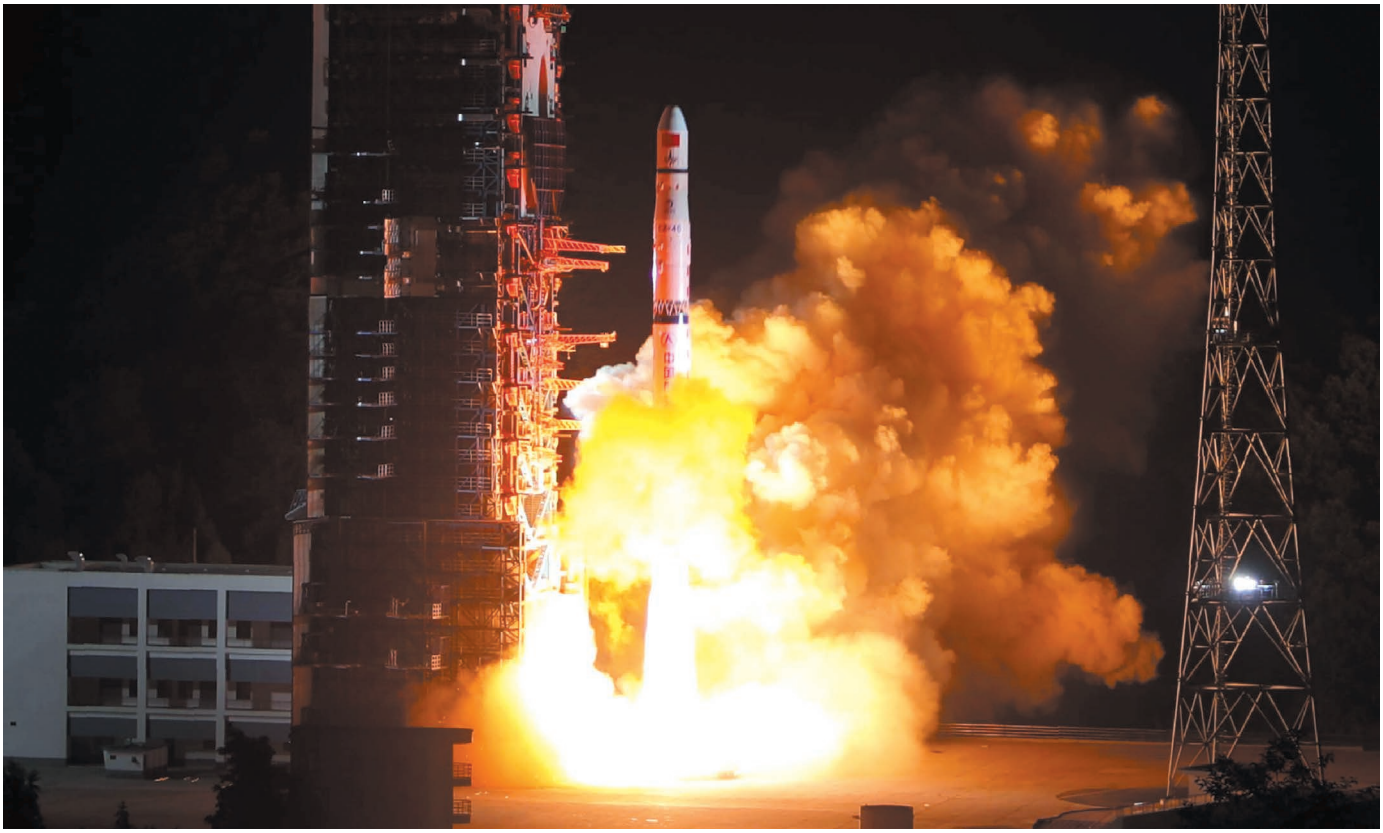
Kilauea probably isn't done erupting yet, if the 1924 event is a guide: it went on for more than 2 weeks and caused more than 50 steam explosions. Hawaii's state volcanologist, Bruce Houghton, says that the current event seems to be a type of complex eruption that occurs only once every 50 to 80 years.

The recent activity has prompted the state of Hawaii to order the evacuation of several areas near Kilauea, protecting residents from the greatest eruption danger — flying rocks. But ash could still present a problem for people with respiratory problems, and officials are monitoring the wind and weather patterns to predict where it will go. Other hazards include plumes of sulfur dioxide and a toxic haze that forms when lava reaches the ocean, as it did on 19 May.

Houghton says that it may be hard to know when it is safe for people to return to evacuated areas. In the 1960s, an eruption at Kilauea stopped for a year and a half before the volcano resumed spewing ash. "It takes quite a long time to be certain things are completely over," he says. "The past record suggests it might go for months or years."

Kilauea's latest eruption "is an exciting event but it's coming at a cost", Poland says, noting that homes have been destroyed and tourism is suffering. "We as the scientific community feel we owe it to the people being impacted to get it right and learn as much as we can," he says. ■

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



AFP/GETTY

The Queqiao spacecraft and two radio-astronomy experiments launched from the Xichang Satellite Launch Centre in western China on 21 May.

SPACE

China shoots for Moon's far side

Queqiao probe carries technologies that could one day explore the Universe's dark ages.

BY DAVIDE CASTELVECCHI

China has taken its first major step in a groundbreaking lunar mission. On 21 May, a probe launched from Xichang Satellite Launch Centre to head beyond the Moon, where it will lie ready to act as a communications station for the Chang'e-4 lunar lander. The nation hopes that the lander will, later this year, become the first craft to touch down on the far side of the Moon.

The relay probe, named Queqiao and designed by the Chinese Academy of Sciences, also carries two pioneering radio-astronomy experiments. Both are proof-of-principle missions designed to test technologies for exploring a period in cosmic history known as the dark ages. These first few hundred million years of the Universe's existence, before galaxies and stars began to form, are all but impossible to study from Earth. But the spectrum of radiation from this age — when matter was distributed nearly uniformly across space as a thin, cold haze — could reveal information about the distribution of ordinary matter compared with dark matter in the Universe.

The first of the two experiments is the Netherlands–China Low-Frequency Explorer (NCLE). It will remain attached to Queqiao, which will linger around 'Earth-Moon L2' — a gravitational resting point about 60,000 kilometres beyond the Moon that tracks the Moon's orbit around Earth. The Dutch-built NCLE experiment will try to exploit the relative quiet there to measure radio waves with frequencies between about 1 megahertz and 80 megahertz, coming from the Solar System, the Galaxy and beyond.

"The experiment is an important first step toward investigating the dark ages and cosmic dawn."

Much of this frequency band is blocked by Earth's atmosphere, but cosmologists expect it to contain information from the dark ages. Around the upper end of this band also fall the 'cosmic-dawn' signals from the first stars, which lit up around 200 million years after the Big Bang, and were apparently detected for the first time in Australia earlier this year. Other experiments are trying to replicate those results — but the NCLE is testing technologies for identifying lower-frequency signatures from the dark ages.

For at least part of its orbit, Queqiao will be eclipsed by the Moon, as seen from Earth, which could benefit the NCLE because its antennas will be further shielded from the radio noise that constantly leaks from our planet. Still, observation time and the bandwidth for sending data back to Earth will be limited. And because Queqiao is designed primarily as a data-relay station (its name comes from a folktale about magpies that form a bridge across the sky), it is not optimized for radio astronomy. That means it will be challenging, if not impossible, for this demonstrator mission to detect the dark-ages signal, says Heino Falcke, a radio astronomer at Radboud University Nijmegen in the Netherlands who is the experiment's science leader. Nonetheless, the NCLE "is pioneering and an important first step toward investigating the dark ages and cosmic dawn", says Jack Burns, an astrophysicist at the University of Colorado Boulder who is leading a proposal for a NASA mission with similar objectives.

To avoid jeopardizing the Queqiao probe, mission control will deploy the NCLE's antennas only after the Chang'e-4 lander's

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

mission is completed, says Marc Klein Wolt, a Radboud astronomer who is NCLE's manager. But the NCLE might go on collecting data for several years, he says.

SATELLITE BREAK-OFF

The second experiment that launched with Queqiao consists of two smaller satellites called Longjiang-1 and Longjiang-2, which will detach from the mothership and orbit the Moon. Built by researchers at the Harbin Institute of Technology in China, the instruments will test technology for a radio-astronomy technique called very-long baseline interferometry (VLBI). This approach combines data from multiple radio antennas to create images of much higher resolution than would be possible with a single dish.

Falcke and others have long studied the possibility of doing VLBI with a large array of lunar orbiters — or on the lunar surface — to map variations across the sky in signals from the dark ages and cosmic dawn. Klein Wolt says that his team might experiment with combining data from NCLE with those from the two lunar orbiters, and even from a radio antenna on the Chang'e-4 lander itself.

The Chang'e-4 mission is another step in China's ambitious lunar-exploration programme, which aims to establish a Moon base in the next decade, and to begin human exploration in the 2030s. The lunar lander will carry a rover and was originally designed as a back-up for Chang'e-3, which in 2013 became the first craft since 1976 to soft-land (rather than crash-land) on the Moon. Chang'e-4 has now been repurposed, and the mission's main scientific goal is to study the geology of the hidden side of the Moon, which is pockmarked with many more small craters than the familiar near side.

The lander carries several experiments, including a sealed ecosystem, built by Chongqing University, which will test whether potato and thale-cress (*Arabidopsis*) seeds sprout and photosynthesize as silkworm eggs hatch and the worms produce carbon dioxide. Another experiment will measure the radiation that will confront future astronauts who visit the lunar surface. The rover, which will separate from the lander to move around the surface of the Moon, will carry instruments such as a solar-wind detector built by a Swedish team. ■

PUBLISHING

Open-access drive spreads in Europe

Negotiators share tactics to broker new deals with publishers.

BY HOLLY ELSE

Bold efforts to push academic publishing towards an open-access model are gaining steam. Negotiators from libraries and university consortia across Europe are sharing tactics on how to broker new kinds of contracts that could see more articles appear outside paywalls. And inspired by the results of a stand-off in Germany, they increasingly declare that if they don't like what publishers offer, they will refuse to pay for journal access at all. On 16 May, a Swedish consortium became the latest to say that it wouldn't renew its contract, with publishing giant Elsevier.

Under the new contracts, termed 'read and publish' deals, libraries still pay subscriptions for access to paywalled articles, but their researchers can also publish under open-access terms so that anyone can read their work for free. Advocates say such agreements could accelerate the progress of the open-access movement. Despite decades of campaigning for papers to be published openly — on the grounds that the fruits of publicly funded research should be available for all to read — scholarly publishing's dominant business model remains to publish articles behind paywalls and collect subscriptions from libraries (see 'Growth of open access'). But if many large library consortia strike read-and-publish deals, the proportion of open-access articles could surge.

"There is a serious ground for change across Europe," says Koen Becking, chief negotiator for the VSNU, a consortium of 14 institutes in the Netherlands. In 2014, the VSNU was the first national group to negotiate a subscription deal that included rights for its scholars to publish all of their work openly. It has since agreed several more that include varying levels of open publishing. Consortia in Austria, the United Kingdom, Sweden and Finland have struck similar deals, and Switzerland will start to negotiate its first open-access contracts this year. A survey by the Brussels-based European University Association, published in April, reported that, last year, 11% of negotiating consortia in Europe made deals that took into account open-access publishing costs, but 63% planned to do so in the future.

On 2 May, negotiators from countries across Europe agreed to align their bargaining strategies at a closed meeting in Berlin attended by the European Commission's special envoy for open access, Robert-Jan Smits. According to Gerard Meijer, one of the German negotiators present, consortia are "frustrated" by the lack of progress in talks and feel the limits of partnerships between institutions and large publishers "have been reached. It is up to us now to act, and to step out of these negotiations if these are going nowhere," he says.

The meeting was the latest in a string of events in which negotiators from different countries swapped tactics. "More and more people are willing to share their experiences," says Matthijs van Otegem, director of the library at Erasmus University in Rotterdam, and chair of the open-access working group at the Association of European Research Libraries (LIBER) in The Hague, the Netherlands.

In September last year, LIBER published a list of principles to guide negotiators seeking to change their deals. These include ending non-disclosure agreements that publishers customarily place on contracts (which would enable negotiators to compare deals in different countries) and not agreeing to price hikes without open-access agreements in place.

A key driver behind the activity in Europe is the European Commission's goal that, by 2020, all research will be freely accessible as soon as it is published. Dutch negotiators have been tasked with brokering a deal that meets ▶

MORE ONLINE

TOP NEWS



LARS BOEHME (SMRU)

Deep-diving seals recruited to measure climate change go.nature.com/2kdlsOh

MORE NEWS

- Share of African American men going into medicine hits historic low go.nature.com/2x1azd3
- Society rescinds award to top seismologist go.nature.com/2lujpfw
- UK's powerful funding body takes shape go.nature.com/2x5dsfW

NATURE PODCAST

The cost of climate change; cleverer cab journeys; and peering through matter with muons nature.com/nature/podcast

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

this vision. In Sweden, negotiators have set themselves the target of complete open-access research by 2026, and in Switzerland the planned date is 2024. But word is spreading outside Europe, too: last month, negotiators in Germany travelled to South Korea to discuss their work with consortia there, and representatives from the University of California system attended the Berlin meeting.

DEAL OR NO DEAL

The situation in Germany has shown that 'no deal' is an option, van Otegem says. Since 2016, a university consortium there has held out on a new deal with Elsevier. Despite the stand-off, the publisher has not stopped German scholars accessing its journals — suggesting that universities need not fear researchers' wrath if negotiations stall. Since then, other consortia have also announced 'no deals' with publishers.

One reason that libraries no longer fear an end to their contracts is that a growing number of free versions of paywalled articles can be found online as preprints or accepted manuscripts, notes Heather Joseph, executive director of the Scholarly Publishing and Academic Resources Coalition (SPARC), an advocacy group in Washington DC. Sci-Hub, a website that illicitly hosts copies of papers and is used by academics around the world, is also a big factor, says Joseph Esposito, a publishing consultant in New York City. "Without Sci-Hub the researchers would be screaming at the libraries and state agencies not to cut them off," he says.

Costs are a major sticking point in the stand-offs. A spokesperson for the Royal Society of Chemistry's publishing arm, for instance, says that its current dispute with the VSNU revolves partly around the difficulty of reaching "an agreement which makes the transition to open access sustainable". (Other publishers approached by *Nature*'s news team declined to comment on details of specific negotiations.)

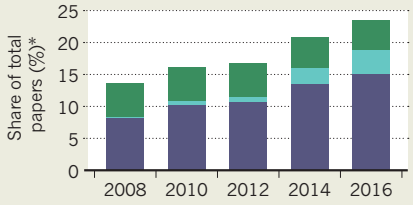
Consortia are generally unwilling to discuss whether the read-and-publish contracts work out to be more expensive, but some say they don't want to agree to contracts that require above-inflation price rises. And in the United States, where open access has less political impetus than in Europe, libraries are trying to save money by cancelling 'big deal' contracts — comprehensive, but expensive, deals for access to large bundles of journals — in favour of à la carte access to the journals their academics use the most. This has happened before, but is now a more common approach, according to a list collated by SPARC.

Steven Inchcoombe, Springer Nature's chief publishing officer, says that some deals that combine reading and publishing costs have been brokered in northern Europe because of strong support for open access from research funders, institutions and governments. But unless more money is available to pay for such deals, they are unlikely to become more popular in the future, he says. "It is in everyone's interest to solve this," he adds. (*Nature*'s news team is editorially independent of its publisher.)

If the stand-offs continue, the VSNU's Becking thinks that negotiators might end up striking deals, but might also simply stop bargaining with particular publishers. In that case, universities could encourage researchers to disseminate their work on alternative platforms, he says. ■

GROWTH OF OPEN ACCESS

In 2016, journals made 18.9% of papers open immediately on publication, up from 11.5% in 2012.



*From Scopus database. †Subscription journals with OA option.

SOURCES: UNIVERSITIES UK (2017)/ M. LAAKSO & B-C. BJÖRK *BMC MED.* **10**, 124 (2012).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



GUT MICROBES JOIN THE FIGHT AGAINST CANCER

The intestinal microbiome seems to influence how well some cancer drugs work. But is the science ripe for clinical trials?

BY GIORGIA GUGLIELMI

Bertrand Routy earned a lamentable reputation with Parisian oncologists in 2015. A doctoral student at the nearby Gustave Roussy cancer centre, Routy had to go from hospital to hospital collecting stool samples from people who had undergone cancer treatments. The doctors were merciless. "They made fun of me," Routy says. "My nickname was Mr Caca."

But the taunting stopped after Routy and his colleagues published evidence that certain gut bacteria seem to boost people's response to treatment[1]. Now, those physicians are eager to analyse faecal samples from their patients in the hope of predicting who is likely to respond to anticancer drugs. "It was an eye-opener for a lot of people who couldn't see the clinical relevance of gut microbes," says Routy, who is now at the University of Montreal Health Centre in Canada.

ILLUSTRATION BY OLA NIEPSUJ

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Cancer has been a late bloomer in the microbiome revolution that has surged through biomedicine. Over the past few decades, scientists have linked the gut's composition of microbes to dozens of seemingly unrelated conditions — from depression to obesity. Cancer has some provocative connections as well: inflammation is a contributing factor to some tumours and a few types of cancer have infectious origins. But with the explosive growth of a new class of drug — cancer immunotherapies — scientists have been taking a closer look at how the gut microbiome might interact with treatment and how these interactions might be harnessed.

After preliminary findings in mice and humans revealed that gut bacteria can sway responses to such drugs, scientists started trying to decipher the mechanisms involved. And researchers are launching a handful of clinical trials that will test whether the gut microbiome can be manipulated to improve outcomes.

Some proponents say that strategies to mould the microbiome could be game-changing in cancer treatment. "It's a smart place to be," says Jennifer Wargo, a surgeon–scientist at MD Anderson Cancer Center in Houston, Texas. But others are worried that the move to the clinic is premature. William Hanage, an epidemiologist at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts, calls the idea "phenomenally interesting", but adds: "I have some anxiety about the notion that only beneficial effects are possible."

INTRIGUING LINK

Although the excitement about microbes and immunotherapy has emerged only in the past three years, some researchers have been exploring connections between gut bacteria and cancer for much longer. Scientists first linked the infectious bacterium *Helicobacter pylori* to gastric cancer back in the 1990s, for example. And since then, other bacteria have been associated with cancer initiation and progression. Some of these microbes activate inflammatory responses and disrupt the mucus layers that protect the body from outside invaders, creating an environment that supports tumour growth. In other cases, they promote cancer survival by making cells resistant to anticancer drugs.

But gut bacteria can also help fight tumours. In 2013, a group led by Laurence Zitvogel[2] at Gustave Roussy and one led by immunologists Romina Goldszmid and Giorgio Trinchieri[3] at the National Cancer Institute in Bethesda, Maryland, showed that some cancer treatments rely on the gut microbiome activating the immune system.

Zitvogel's team found that the chemotherapy drug cyclophosphamide damages the mucus layer that lines the intestine, allowing some gut bacteria to travel into the lymph nodes and spleen, where they activate specific immune cells. For mice raised without microbes in their guts or given antibiotics, the drug largely lost its anticancer effects.

Following this observation, Zitvogel decided to explore whether bacteria in the gut might influence responses to a class of immunotherapy drugs called checkpoint inhibitors. These drugs, typically antibodies to cell-surface molecules such as CTLA4 and PD1, unleash a person's immune system against tumour cells, and are used to treat several types of cancer (see 'A little help from their friends'). But only 20–40% of people respond to treatment[4].

In 2015, Zitvogel and her team showed that microbe-free mice failed to respond to one such drug, and mice given a particular bacterium, *Bacteroides fragilis*, responded better than did mice without it[5].

The idea started to spread. Thomas Gajewski, a cancer clinician at the University of Chicago in Illinois, reported that *Bifidobacterium* microbes increased the response to cancer immunotherapy in mice[6]. These gut-dwelling bacteria acted by boosting the ability of some immune cells to initiate a response against tumours.

Wargo saw these results presented at a meeting in 2014, and on returning to Texas, immediately started to collect stool samples from people with skin cancer who were about to undergo immunotherapy at her institution. Last November, Wargo[7], Gajewski[8] and Zitvogel[1] all published results in *Science* linking positive immunotherapy responses in people to specific varieties of gut bacteria. The samples that Routy had collected in Paris helped Zitvogel's team to also show that people who had taken antibiotics for unrelated infections tended to respond poorly to immunotherapy.

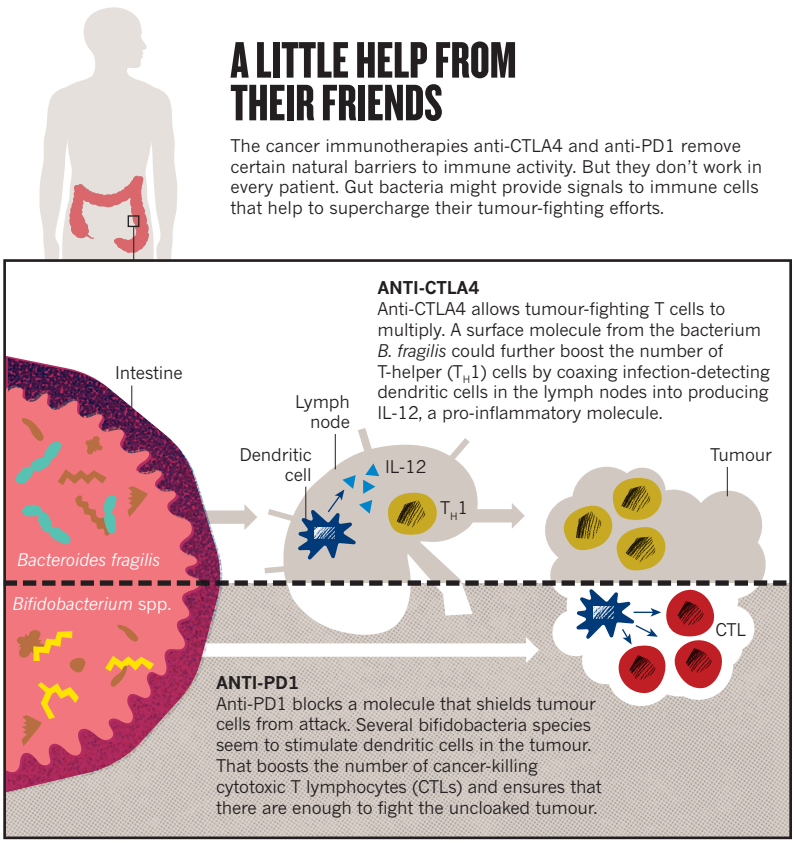
To solidify the relationships, the researchers transferred bacteria from the human participants into the intestines of mice with comparable cancers. Rodents who got 'beneficial' bacteria developed smaller tumours than did mice that received microbes from people who hadn't responded to treatment. "All of this work has been very exciting," says Neeraj Surana, a microbiologist at Boston Children's Hospital. "They've opened up the possibility for a clear therapeutic application of microbiome science."

HEADING TO THE CLINIC

Researchers are now running with that possibility. Hassane Zarour, an immunologist at the University of Pittsburgh in Pennsylvania, partnered with the global pharmaceutical company Merck to collect faecal bacteria from people who respond to treatment with a checkpoint inhibitor and transfer them into the intestine of non-responders, a process called faecal microbiome transplant. Merck has invested about US$900,000 into this trial, which is set to start in the next few weeks.

Wargo is planning a similar trial. Together with the Parker Institute for Cancer Immunotherapy in San

"THEY MADE FUN OF ME, MY NICKNAME WAS MR CACA."



© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

"IT WAS AN EYE-OPENER FOR A LOT OF PEOPLE WHO COULDN'T SEE THE CLINICAL RELEVANCE OF GUT MICROBES."

Francisco, California, and the biotech company Seres Therapeutics in Cambridge, Massachusetts, she expects to test whether faecal transplants can reshape the gut microbiome of non-responders in a beneficial way.

These microbiome transplants are becoming a mainstream treatment for some non-cancer illnesses. In February, for example, the Infectious Diseases Society of America recommended that physicians use these procedures to treat people with bowel infections caused by the bacterium *Clostridium difficile* who had failed to respond to other treatments. But the approach has downsides. To avoid the risk of inadvertently infecting people with pathogenic microbes, researchers must be careful in how they select donors and screen faecal material before transferring it to recipients. That's why, in addition to faecal transplants, Seres Therapeutics, the Parker Institute and Wargo will test a pill containing a set of spore-forming bacteria that have been purified from the faeces of responding patients.

Gajewski and his partners at Evelo Biosciences, a biotech company in Cambridge, are using a similar approach. Their trial will assess the effects of two pills containing single bacterial strains in people with different types of cancer, including colon and skin cancer.

Zitvogel is not planning to start clinical trials but she has co-founded the Delaware-based start-up EverImmune, which is developing a microbiome-based pill.

It's still unclear exactly how microbes might interact with immunotherapeutics. A widely accepted hypothesis is that some boost the body's response against tumours by regulating how easy it is to activate the immune system. But the precise mechanisms, including which bacteria modulate which immune cells, remain a mystery.

The researchers hope that the clinical trials will help to clarify things. Wargo, for instance, is exploring bacterial metabolites. Her team hopes to find specific metabolic signatures of a good outcome in the stools and blood of people who respond to therapy, as well as to document the numbers of immune cells in the blood and tumours of trial participants.

Gajewski suggests that microbes might be unleashing the immune response by stimulating the gut cells to produce certain molecules. His team is testing whether circulating immune-cell precursors change their behaviour when specific bacteria are given to mice. At the same time, the group is trying to pin down which species might be driving the positive outcomes.

TOO EARLY, OR JUST RIGHT?

Given the uncertainties, some scientists argue that testing these approaches in humans is risky. Some trial participants could experience side effects, Surana says. And changing the make-up of an individual's microbiome might predispose them to other health problems.

Faecal transplants come with a lot of unknowns. They have proved safe and effective in many people without cancer, Wargo says, but they have also been associated with unexpected effects, including one case in which the procedure led to weight gain and obesity[9]. "Should we look for safety signals on these trials? Absolutely." Wargo says, "But I strongly feel that we need to go into these trials. We need to design them well. We need to really learn from these trials."

Gajewski, who plans to test the effects of just one bifidobacterial strain at a time, says there's good reason to be confident. "People have eaten bifidobacteria for a thousand years," he says. The bacteria are present in the gut of infants and decline in number as the people grow up, so they should at least be safe, he adds.

But it's unclear whether a single species can help people with cancer and, if so, what bacterium that is. The papers published in *Science* last year all associated different bacteria with the best outcomes, even for the same cancer and therapy.

The researchers looked at people with cancer from France and the United States, so diet could account for some of the differences, Wargo says. But variations in sample collection, data analysis and statistical methods could also have skewed the results, says Joël Doré, a biologist at the French National Institute for Agricultural Research in Paris who in 2011 helped to launch the International Human Microbiome Standards (IHMS) project with the aim of improving data reproducibility in microbiome research.

Hanage says that even the two studies[7,8] that analysed people in the United States with the same type of cancer identified only a partially overlapping set of microbes associated with positive outcomes. If researchers don't work out the reason for these differences, they might not be able to interpret the outcomes of the trials, Hanage says.

Before starting clinical trials, the three groups should try to reproduce each other's results and converge on a set of 'beneficial' microorganisms, Hanage argues. "Any of these bacteria could be a useful approach." But inconsistencies might mean that the results are not reproducible.

It's a concern common to microbiome research. "A lot of findings have proven to either not stand up or be considerably more complicated than they first appeared," Hanage says. Standards such as those developed by the IHMS project should help, but scientists will be reluctant to take them on board, says Susan Erdman, a microbiologist and cancer biologist at the Massachusetts Institute of Technology in Cambridge. Doing so would come at the cost of innovation, she argues — it's by experimenting in different settings that researchers make discoveries.

Wargo says that the community should standardize its approaches for collecting samples and doing analyses, as well as for validating studies in larger groups of patients. Since last year, her group has analysed stools from more than 500 people with skin cancer who had received different therapies. In parallel with the Paris-based team led by Zitvogel, the researchers are analysing patients treated with two combined immunotherapies to work out which gut bacteria mediate a response to that combination. Wargo hopes that the gut microbiome could eventually help to identify which patients will respond to which anticancer treatments. "Can we use it as a biomarker? It's a provocative question," she says.

In the short term, there will be a whole lot more sample collection. And this time around, it's likely that fewer oncologists will raise an eyebrow, says Routy, who is now investigating how the gut microbiome boosts immunotherapy with his own group. In cancer therapy, "gut microbes have gone from ignored to super-popular organisms", he says. Now, they'll just have to live up to their reputation. ■

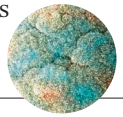
Giorgia Guglielmi *is an intern with* Nature *in Washington DC.*

1. Routy, B. *et al. Science* **359**, 91–97 (2018).
2. Viaud, S. *et al. Science* **342**, 971–976 (2013).
3. Iida, N. *et al. Science* **342**, 967–970 (2013).
4. Sharma, P., Hu-Lieskovan, S., Wargo, J. A. & Ribas, A. *Cell* **168**, 707–723 (2017).
5. Vétizou, M. *et al. Science* **350**, 1079–1084 (2015).
6. Sivan, A. *et al. Science* **350**, 1084–1089 (2015).
7. Gopalakrishnan, V. *et al. Science* **359**, 97–103 (2018).
8. Matson, V. *et al. Science* **359**, 104–108 (2018).
9. Alang, N. & Kelly, C. R. *Open Forum Infect. Dis.* **2**, ofv004 (2015).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

COMMENT

GENETICS Carl Zimmer's new history of heredity **p.489**



ARTIFICIAL INTELLIGENCE Are algorithmic works by robot artists or artists' robots? **p.490**

CONSERVATION A call to involve young people and locals in coral-reef protection **p.492**

TECHNOLOGY Lithium mining in Chile's Atacama puts people and ecosystems at risk **p.492**

SANJIT DAS/PANOS PICTURES



Women in Odisha, India, where in 2014 the Dongria Kondh forest tribe won a lawsuit to stop a bauxite mine from opening.

The global south is rich in sustainability lessons

Educators must share how communities in the developing world manage environmental change — a Western bias limits progress, argues **Harini Nagendra**.

In a Bangalore slum, Dhanalakshmi tends six plant pots balanced on a wall. They contain shoots of holy basil (or tulasi, *Ocimum tenuiflorum*). I asked her why she does this, in a cramped space with an unreliable water supply. She told me that the plants replace her tiny roadside kitchen garden, which she lost when the street was widened. The wind blew the basil seeds into the pots. "How can one turn away a guest, even if they come uninvited?" she said.

Dhanalakshmi's deep, personal connection to nature shapes her actions, even though she lives far from the countryside. Such attachments are shared by many people around the world. They run through centuries of Indian thinking on sustainability: nature offers material benefits; it is part of people's cultural identities and often viewed as sacred. Protecting nature also confers social merit. A stone inscription from AD 1340 describes the motivation of Chenneya Nayaka, the ruler of a region near Bangalore, for building an irrigation tank: 'to support animals, cattle, birds, and all other living beings, and the service at all times of (the goddess of water) Ganga Devi'[1].

In the early twentieth century, Mahatma Gandhi fought poverty and injustice through peaceful civil resistance. He championed local production, education, health care and self-sufficiency. Inspired by Gandhi's ideas, members of the Chipko and Appiko environmental movements hugged trees in the 1970s and 1980s to prevent them ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

from being felled[2]. In 2014, after 12 years of campaigning, the Dongria Kondh forest tribe in India's Odisha state won a lawsuit to stop a bauxite mine from opening and ruining the hillsides that they revered and depended on for food.

These strongly rooted local movements have brought sustainability issues into everyday conversations in India. They have inspired generations of activists. Yet most university courses on sustainability omit them. Teachings still have a Western focus, even in India. Most books on sustainability frame the discourse in terms of Earth's finite resources and rising population.

The limited Western view of sustainability is stifling progress, just as the world faces crises over water, climate change, energy and biodiversity. That view also does a disservice to the variety and creativity of thinking and actions on sustainability in societies across the globe. Developing countries face the most acute challenges in this regard, yet they have the widest gaps in knowledge. Solutions that work in one place might fail in another. Excessive consumption, inequity and social injustice are not questioned enough.

"Excessive consumption, inequity and social injustice are not questioned enough."

At Azim Premji University in Bangalore, my colleagues and I see sustainability differently. We have moved away from framing it exclusively around limits to growth and conserving natural resources. Instead, we emphasize the connections between communities, ecosystems and social justice. In an online course, for instance, we discuss the '3 Fs' — finitude (or limits), fragility and fairness (see go.nature.com/2t3rfdd). As well as university students, from undergraduate to postgraduate level, we teach bureaucrats, educators, corporate executives and practitioners through online courses and in-class curricula.

Sustainability education must be more globally inclusive. Only then can the discipline deliver the transformative change the world needs, rather than tinkering with business-as-usual.

LOCAL DIFFERENCES

Sustainability is usually discussed as if 'one size fits all'. Calls to action target the individual: plant a tree, ride a bike, compost your food scraps (see go.nature.com/2rwsupi). Or they focus on markets and corporations: invest in renewable energy and green buildings. On these terms, local contexts are irrelevant and materials matter more than people — buying an electric car is 'green', even if the cobalt for its battery might have come from a small-scale artisanal mine with horrific labour conditions.



Most of the sustainability workshops I've attended in the past three years in India focused on reducing resource use. We discussed emissions cuts, recycling and the circular economy while sitting in air-conditioned rooms in luxury hotels, sipping bottled water. Poverty, environmental justice and governance were not mentioned.

Most 'classic' writings on sustainability present people as the problem, not as a collective source of strength. These include Garrett Hardin's 1968 essay *The Tragedy of the Commons*, Paul Ehrlich's 1968 book *The Population Bomb* (Sierra Club/Ballantine Books) and *The Limits to Growth*, a 1972 report by Donella Meadows and colleagues. For example, *The Population Bomb* opens by describing the streets of Delhi in 1966 as "alive with people" — eating, washing, sleeping, visiting, arguing, screaming, urinating, clinging to buses and herding animals. An obsessive focus on overpopulation has led to millions of forced sterilizations worldwide[3].

Much is left out from these accounts. Political economist Elinor Ostrom, in her influential work on the commons[4], demonstrated the powerful capacity of people when they are organized in collectives. Discussions of increasing consumption are largely absent from classic writings on sustainability, despite the cases for sufficiency made by twentieth-century Indian thinkers. German economist Ernst Friedrich Schumacher, whose book *Small Is Beautiful* (Blond & Briggs, 1973) championed local, sustainable technologies that empower people, was inspired by Gandhi and the Indian economist J. C. Kumarappa.

In our courses, we shine a light on these issues. And we show our students how iconic local movements in the global south have been just as influential in their regions as US environmentalists such as John Muir, Aldo Leopold and Rachel Carson were in theirs. In Latin America, for example, the concept of *buen vivir* (living well) has widespread resonance. It espouses harmony with

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



SANJEEV GUPTA/EPA/REX/SHUTTERSTOCK

Protesters in Bhopal march against proposed dams on India's Narmada River.

nature, prioritizing community well-being and respecting plurality of thought[5]. From Nepal to Mexico, indigenous communities have managed forests and prevented poaching in locally run reserves. Those fenced off by governments have fared much worse.

We dissect lots of case studies in our courses and workshops. One clear lesson is that transplanted solutions often backfire. For example, a ban on livestock grazing degraded the ecosystem of the Keoladeo National Park — a region of wetlands in Rajasthan that is rich in bird life. The Indian government introduced the policy, following US practices, to protect supposedly pristine landscapes from trampling. Cattle were duly ejected from Keoladeo in 1982. But the diversity of birds and other wildlife plummeted — the canals became clogged with weeds and grasses, which were previously eaten by the cattle[6].

Community relations are important. In 1974 in southern India, the indigenous Soliga tribe was banned by the government from setting controlled fires in the forests of the Biligiri Rangaswamy Temple Tiger Reserve. The community warned policymakers of the consequences, but they did not listen. Without control through burning, the invasive shrub *Lantana camara* ran riot, choking vegetation and reducing fodder and food[7]. A well-meaning policy, influenced by ecological ideas about succession, has ended up damaging the forest as well as relationships between the tribe and the forest department.

Technologies, too, can bring more harm than good. In the push for solar energy, land has been acquired forcibly from poor farmers. In Nepal and India, replacing traditional mud-lined irrigation tanks and channels managed by farmers with centrally managed cement-lined canals has increased maintenance costs and damaged social capital. The canals silt up, and farmers no longer meet and work together to repair them.

In all these cases, a science-focused administration ignored the knowledge of local communities. Officials failed to appreciate the fragile social and ecological interconnections in these densely inhabited, biodiverse landscapes.

Our students tell us that some of these findings come as a surprise. The fragility of communities must be considered alongside ecosystems. No technology, no matter how good, is a magic silver bullet. Narratives such as Dhanalakshmi's can contradict widely held assumptions — for instance, that strong ties to nature exist only in pre-industrialized societies.

Some students have reversed their opinions. For instance, those who initially supported diverting water from full rivers into others that ran dry now understand the importance of maintaining river-basin flows for ecosystems and communities. Practitioners have been sufficiently inspired by these examples of ecosystem management through fire and grazing to begin altering their approaches to working with communities.

GRASS-ROOTS CAMPAIGNS

There are hundreds of environmental movements across Asia, Africa and Latin America. Although they are diverse, they have features in common. Such movements emphasize environmental justice and tend to emanate from local cultures. They are often led by women's collectives, and use non-violent means of protest. They question the industrial route to development and champion collective action. Social justice and the rights of nature are given the same prominence as limited resources. Multigenerational thinking often features.

For example, in 2008, catalysed by a coalition of indigenous groups, Ecuador became the first nation to incorporate rights of nature in its constitution[5]. Bolivia followed suit in 2009. These rights include protection, restoration and respect for existence.

Grass-roots campaigns can be powerful in questioning unsustainable paradigms and changing minds, even when they don't prevail. For 33 years, the Narmada Bachao Andolan mass social movement has marched and brought court cases to stall dam construction on India's Narmada River, which runs from Madhya Pradesh to the Arabian Sea. It did not stop the dams, but it has raised awareness of the consequences for people and places of big, top-down developmental projects.

Local groups tackle a wide range of issues. In Indian cities, groups such as Hasiru Dala work with poor rag pickers to collect and recycle waste. In Bangalore, residents are liaising with municipal authorities, cattle grazers and fishers to restore the

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



ANDRESSA ANHOLETE/AFP/GETTY

Indigenous people protesting in Brasilia over the government's failure to safeguard their land.

Kaikondrahalli and Jakkur lakes. Across rural India, community groups collaborate to protect local forests, create sustainable jobs and provide incomes while protecting biodiversity. Agro-ecological initiatives such as the Deccan Development Society in Medak district, the Foundation for Ecological Security in Anand and the Timbaktu Collective in Anantapur district work with farmers to restore forests and common land, and to promote organic methods of farming and soil-friendly crops such as pulses and millets.

To encourage networking and dialogues, the Vikalp Sangam or Alternatives Confluence, a non-profit collaborative discussion forum and website, documents grass-roots experiences (www.vikalpsangam.org). Governments can help to scale up such initiatives. For instance, a programme of the state government of Kerala called Kudumbashree works with women's groups on empowerment, livelihoods and sustainability.

Governments and grass-roots initiatives cannot solve all sustainability issues in isolation, especially in a country such as India that is accelerating towards an industrialized and urban future. We just have to look outside our classroom windows to see the negative impacts of India's relentless growth.

But sustainability and conservation are dismal disciplines. The next generation needs cases of hope to counter narratives of gloom and doom. And they need to know that successes can be found on their doorstep, not just in the West.

GLOBAL LESSONS

Sustainability curricula must be rethought. It is important to learn about, teach and communicate ways to reduce resource consumption. It is even more crucial — and much harder — to transform world views and dismantle unsustainable paradigms of development and growth.

Sustainability needs to be defined as encompassing natural resource conservation as well as social justice and collective action. Such world views must go beyond purely utilitarian concepts of nature. Educators should tailor their lessons to be more globally inclusive. The United Nations Environment Programme, the UN Educational, Scientific and Cultural Organization and Future Earth, a global research hub for sustainability, should develop more diverse educational and outreach materials. Academics, indigenous and local communities, practitioners and environmental activists must all be involved.

> ***"Researchers from the global south should lead study projects, and funding agencies should provide support."***

To expand the range of cases examined, researchers must document contemporary grass-roots attempts to engage with sustainability in the global south, such as those discussed here. A good start is the Seeds of a Good Anthropocene project run by Future Earth. This shares case studies and tools to support positive visions of futures that are socially and ecologically desirable, just and sustainable. More examples should be drawn from Africa, Latin America and Asia. Researchers from the global south should lead study projects, and funding agencies should provide financial support at the scale needed to maintain such leadership.

Key topics to explore include how communities reshape traditional approaches to grapple with twenty-first-century challenges, how they address gender and caste inequities, and how philosophies and faiths influence people's attitudes to nature. For example, Kudumbashree, the Foundation for Ecological Security and the Deccan Development Society are enabling women farmers, fishers and grazers to take the lead in public decision-making. The Dongria Kondh tribe's belief that each component of the landscape has sacred significance shaped its rejection of commercial mining[8].

Sustainability curricula cannot rest on just-so stories. A set of universal principles needs to be derived, while respecting local contexts. Ostrom's framework[4] for governing shared resources is a good basis. Local crafting of rules, limiting free riders through monitoring, and strong local leadership are used in such disparate cases as community forests in Nepal, Subak irrigation systems in Bali's rice fields, alpine grazing commons in Switzerland and Satoyama agricultural landscapes in Japan.

Science and technology can only go so far. Without understanding alternative imaginations — such as the cosmology of the Dongria Kondh or the compassion of Dhanalakshmi — we limit our power to effect change. ■

Harini Nagendra *is professor of sustainability at Azim Premji University, Electronic City, Bangalore, India.*
e-mail: harini.nagendra@apu.edu.in

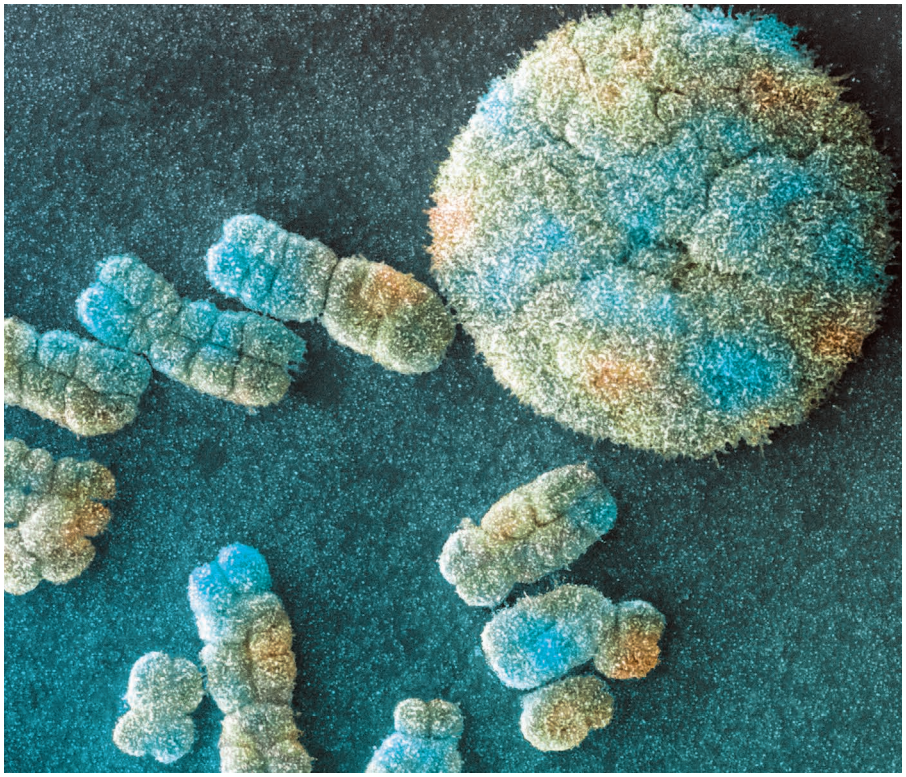
1. Nagendra, H. *Nature in the City: Bengaluru in the Past, Present and Future* (Oxford Univ. Press, 2016).
2. Guha, R. *How Much Should a Person Consume? Environmentalism in India and the United States* (Univ. California Press, 2006).
3. Connelly, M. *Fatal Misconception: The Struggle to Control World Population* (Harvard Univ. Press, 2008).
4. Ostrom, E. *Governing the Commons* (Cambridge Univ. Press, 2015).
5. Gudynas, E. *Development* **54**, 441–447 (2011).
6. Lewis, M. *Conserv. Soc.* **1**, 1–21 (2003).
7. Madegowda, C. *Econ. Polit. Wkly* **44**, 65–69 (2009).
8. Tatpati, M., Kothari, A. & Mishra, R. *The Niyamgiri Story: Challenging the Idea of Growth without Limits?* (Kalpavriksh, 2016); available at https://go.nature.com/2k9ju8z

CORRECTION

In the Comment 'Cybersecurity needs women' (*Nature* **555**, 577–580; 2018), the photo of female programmers was captioned incorrectly. They were at the US Army Ballistics Research Laboratory in 1962, not working on ENIAC at the University of Pennsylvania in the 1940s. Also, the figure of 57% cited for women in the US workforce was actually for women in the US professional workforce.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

POWER AND SYRED/SPL



Human chromosomes and a nucleus in a false-colour image taken by scanning electron microscope.

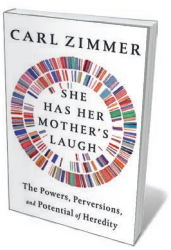
HEREDITY

Beyond the gene

Nick Lane relishes Carl Zimmer's history of heredity in all its messiness, from genes and culture to epigenetics.

Few subjects have afforded more room for doubt, or caused more harm through false certainty, than heredity. In *She Has Her Mother's Laugh*, an illuminating survey of the concept through history, science writer Carl Zimmer shows that scientists have often clung to travesties of the truth — and that we are still in danger of doing so.

The book is a beguiling narrative of more than 600 pages. It blends popular science and history with a personal journey, culminating in a plea for a nuanced view of heredity. Zimmer ably navigates some of the most fraught developments in research, politics, religion and race: from eugenics, slavery and genocide to IQ and genetic engineering in humans. He combines a deep personal empathy with clear scientific understanding. For instance, in presenting controversial figures such as Henry Goddard — who coined the term 'moron' and helped to foster the US eugenics movement in the early twentieth century — he examines their hopes, fears and delusions, before dispassionately gutting their scientific errors and the disastrous consequences.



She Has Her Mother's Laugh: The Powers, Perversions, and Potential of Heredity
CARL ZIMMER
Dutton (2018)

Compellingly, Zimmer delves into his own genome. After having it sequenced at 90% coverage by Illumina in San Diego, California, he got his hands on the raw data, and approached experts such as Dina Zielinski of the New York Genome Center to help him unravel his genes' secrets. Zimmer uses this backstory to illustrate how genomes break up into millions of short stretches of DNA, each with its own history from around the world.

Being told you have ancestors everywhere is one thing; it's quite another to pin that down with visceral intensity. Of the 3,559,137 bases in Zimmer's genome that differ from the human reference (a representative sequence based on a number of donors) he shares 1.4 million single-nucleotide polymorphisms, or SNPs, with two volunteers from China and Nigeria — plus another 530,000 with the Chinese individual and 440,000 with the Nigerian. On top of his roughly 1% Neanderthal inheritance (standard for a person of European descent), Zimmer even has a few Denisovan genes. We should think of the Denisovans as the eastern Neanderthals, he explains. One of their genes, *EPAS1*, might even have helped Tibetans to adapt to high altitudes, although most Denisovan DNA dwindled, leaving little more than a hint of our species's promiscuous past.

At a deeper level, the book is a serious treatise on why we need to overhaul our views on heredity. Zimmer shows how the idea evolved from medieval times, with the passing down of possessions, to our modern focus on genes. He recounts how nineteenth-century genetics pioneers Gregor Mendel and August Weismann seemed to bring clarity by defining simple laws of inheritance in sexual organisms, and by distinguishing between sex cells in the germ line and cells in the rest of the body (see J. Maienschein *Nature* **522**, 31–32; 2015). But heredity soon returned to a swamp of ambiguity. Charles Darwin's cousin Francis Galton, a deeply flawed Victorian statistician and racist (who in 1904 founded what would become the Galton Laboratory at University College London; see go.nature.com/2i6uelm) crops up repeatedly, each time with a new layer of nuance or downright murkiness.

It took the best part of a century for Mendelian genetics to be fully reconciled with complex hereditary traits such as height. Sophisticated statistical methods reveal such traits to be 'omnigenic', influenced by millions of genetic markers. Intelligence is even worse; fairly heritable, certainly, but with a complexity that mocks simple ideas of Mendelian inheritance.

The book goes on to tackle meiotic drive, in which 'selfish' genes evade Mendel's laws by killing the 50% of sex cells that lack the selfish elements, so almost all the offspring inherit the selfish genes. Then we're onto cell lineages, where mutations acquired during development make genetic mosaics of us all; and microchimaeras, in which cells slip, in both directions, across the placental barrier between mother and fetus, sometimes persisting for decades and colonizing whole tissues. (The entire lobe of one woman's liver, Zimmer notes, was composed of Y-chromosome-bearing cells from a male fetus, the paternity of which could be traced to her boyfriend.)

Zimmer explicates transmissible tumours, which infect species including dogs and Tasmanian devils, and can persist in populations for thousands of years, picking up new mitochondria from their hosts. He treats transgenerational epigenetic inheritance with due care, showing how some genetic settings controlled by chemical changes can be passed on with the genes themselves, ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

modulating their activity over multiple generations. That can be seen in eighteenth-century taxonomist Carl Linnaeus's 'monstrous' peloria, a toadflax (*Linaria vulgaris*) with unusual, trumpet-shaped flowers — "no less remarkable than if a cow were to give birth to a calf with a wolf's head", as he put it.

Zimmer completes his tour with chapters on the microbiome (some of which is as heritable as anxiety, and partly accounts for the inheritance of traits including weight), and cultural inheritance. Genes are expressed in a human-altered environment, Zimmer notes, and their effects are as plastic as the culture that shapes their selection, right down to social inequalities. Our inherited environment governs our future more rigidly than our genes.

In this encompassing view of heredity, we get a correspondingly nuanced vision of what, for example, germline editing using CRISPR will really mean. By acknowledging the ambiguous way in which genes actually work, and by embracing all these other factors that shape our lives, we make CRISPR less threatening because it is less definitive.

Zimmer deconstructs the idea of the body as a genetic temple, built on Mendel's sacrosanct 'laws', along with genetic determinism. Instead, he calls for a view that includes "culture, epigenetic marks, hitchhiking microbes, or channels we don't even know about yet". His argument is balanced and fair, comprehensive and bang up to date. Whatever your views on the power of genes versus other forms of heredity, you will be in for a few surprises.

There are some weaknesses. Zimmer makes no real attempt to explain how Mendel's laws arose in our single-celled ancestors, and offers rather cursory descriptions of early evolution. And his sympathy for the underdog can go too far. His portrait of crystallographer Rosalind Franklin, for example, seemed to me too partial. You would never imagine, from Zimmer's depiction of her meticulous science, that Franklin had circulated an obituary of the DNA helix nine months before Francis Crick and James Watson's paper on the double helix appeared in *Nature*. But these are quibbles.

In *She Has Her Mother's Laugh*, Zimmer has built a subtle, multifaceted and deep understanding of heredity, grounded in revelatory insights from genome sequencing. And he shows that we will need it to face our uncertain future. ■

Nick Lane *is professor of evolutionary biochemistry at University College London, and author of four books on evolution, most recently* The Vital Question: Why is Life the Way it Is?
e-mail: nick.lane@ucl.ac.uk



***Untitled 2018*, one of Peter Kogler's installations of computer-assisted art.**

ALDO PAREDES FOR RMN-GRAND PALAIS 2018

TECHNOLOGY

Can robots make art?

Laura Spinney encounters a Paris exhibition that probes the concept of algorithmic creativity.

Visiting the exhibition Artists and Robots at the Grand Palais in Paris, I happened on the artist ORLAN, best known for her work involving body modification. She was standing close to her 2017 work *ORLAN and the ORLANOID*, in which her video presence interrogates a lookalike robot on matters of life and death. Having borrowed the robot's lensless glasses for a photo shoot, she needed her own back. I was struck by the robot's lack of reaction as she made the swap. It underscored my answer to the question posed by this exhibition: can a robot create a work of visual art?

My feeling is no, for the simple reason that it can't see. I recommend the show, nonetheless. It forced me to examine what I mean by seeing or — more broadly — sensing the world, and hence what I mean by art.

"Artificial imagination has yet to get off the starting blocks."

Artists and Robots showcases robots and their output in roughly chronological sections. The first recounts how, starting in the 1950s, visionary artists such as Jean Tinguely and Nicolas Schöffer built robots — to begin with, no more than collections of mobile parts driven by motors — to create kinetic art. The second tracks that impulse forward from the digital revolution, starting in the 1970s.

Artists and Robots
Grand Palais, Paris. Until 9 July 2018.

And the third — optimistically entitled 'The robot emancipates itself' — explores their present status and looks to the future.

When robots were all jointed arms and motors, they executed an artist's vision channelled by their own capacities as machines. Modern French artist Patrick Tresset's ironic spin on this relationship features in the first section of the show. In the installation *Human Study #2*, three sets of robot arms and cameras — the 'hand' and 'eye' — repeatedly draw a set of objects including a stuffed fox and a human skull. They are programmed to copy both the objects and Tresset's drawing technique, while introducing small variations that he characterizes as artistic, expressive and obsessional. It's through such serendipitous additions and mistakes, the artist seems to suggest, that the greats became great.

The digital revolution ushered in software and algorithms as artists' tools or assistants, and the possibilities exploded. We see this in stunning works in the second section, from conceptual artist Joan Fontcuberta's self-described "hallucinatory" landscapes to labyrinthine wallpaper from multimedia whizz Peter Kogler. This covers an entire room, so that we seem enclosed in an optical illusion. For me, the works' technical sophistication seems only to accentuate their soullessness,

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

and never more so than when they show up the fallibility of human perception.

In 2003–04 paintings from his Orogenesis series, for example, Fontcuberta takes algorithms that create 3D landscapes from 2D map coordinates, and forces them to re-interpret the landscape paintings of artists such as J. M. W. Turner and Paul Cézanne. The results are highly naturalistic, but they reminded me of Swiss writer Charles-Ferdinand Ramuz's idea that once the human element is banished from a place, it becomes a non-place.

Likewise Michael Hansmeyer's 2017 *Astana Columns*. These architectural forms, created by an algorithm that applies evolutionary principles to repeatedly subdivide a Doric column, were assembled from laser-cut cardboard and other materials. They provoke awe through their sheer complexity, but in the way that a termite mound does: what's impressive is not that they were imagined, but that they were unimagined.

By the time I reached the future-facing section, where I bumped into ORLAN, I had concluded that, notwithstanding the section's title, the robot had not emancipated itself. Pascal Haudressy's 2009 animation *Brain*, for example, evolves thanks to glitches the artist introduced into the governing algorithm, that force the computer to continuously recalculate the coordinates of each pixel. Ultimately, however, it is less impressive than animations of the actual evolution of the human brain.

Although artificial intelligence has advanced in leaps and bounds since the 1950s, artificial imagination has yet to get off the starting blocks. As curators Laurence Bertrand Dorléac and Jérôme Neutres suggest in an explanatory video, these are artists' robots rather than robot artists. But if Artists and Robots doesn't tell you what art is, it does venture into fascinating new territory to tell you what it's not: random copying errors might be necessary, but they are also insufficient.

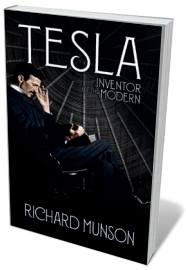
That said, perhaps robot imaginations have already liberated themselves outside the confines of human artists' studios, and their art is radically different from ours; so different that we don't recognize it when we see it, glasses or no glasses. I can't wait for the first exhibition curated by robots — assuming it's advertised to non-robots. ■

Laura Spinney *is a writer and science journalist based in Paris.*
e-mail: lfspinney@gmail.com

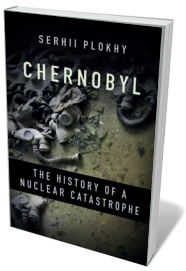
CORRECTION

In the article 'Feynman at 100' (*Nature* **557**, 164–165; 2018), a picture caption mistakenly referred to Richard Feynman lecturing at the California Institute of Technology; the picture was actually taken at California State University, Long Beach.

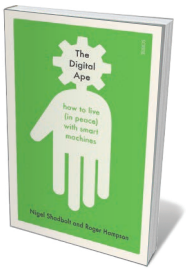
Books in brief



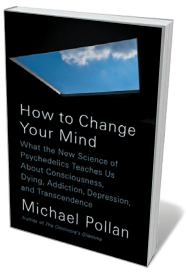
Tesla: Inventor of the Modern
Richard Munson W. W. Norton *(2018)*
Around the turn of the twentieth century, Serbian-born visionary Nikola Tesla authored transformative inventions from the alternating-current system to remote control, drones and (arguably) radio. He even foresaw radar, mobile phones and the Internet. Yet, as Richard Munson reveals in this penetrating biography, Tesla's lack of business sense allowed others to prevail. Munson makes vivid the genius's eventful life, from his mother's inspirational labour-saving inventions to his psychological complexity — and his estimable belief that "technology should transcend the marketplace".



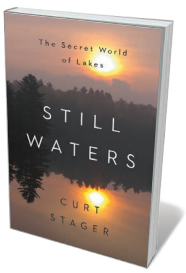
Chernobyl: The History of a Nuclear Catastrophe
Serhii Plokhy Basic *(2018)*
Soon after midnight on 26 April 1986, a turbine test at Ukraine's Chernobyl nuclear power plant went stupendously wrong. The explosion released 14 exabecquerels of radiation; the fallout contaminated 20% of neighbouring Belarus and crossed more than half of Europe. Historian Serhii Plokhy's deft, richly detailed account draws on newly opened archives and weaves in stories of players such as Chernobyl director Viktor Briukhanov. The disaster's roots, he asserts, were a toxic tangle of shoddy construction, human error, flawed governance and complacency in the Soviet nuclear industry.



The Digital Ape
Nigel Shadbolt and Roger Hampson Scribe *(2018)*
Numbed by dire warnings of technological Armageddon? Computer scientist Nigel Shadbolt and economist Roger Hampson dispel the miasma with this superb survey of the landscape we "digital apes" have wrought. Humanity's tool use, spanning everything from hand-axes to CRISPR, has spawned marvels such as a hyper-networked society, "social machines" like Wikipedia and artificial intelligence. But to avoid succumbing to the inherent dangers, Shadbolt and Hampson urge wise choices: to hold Silicon Valley to account, ensure transparent algorithmic decision-making and own our own data.



How to Change Your Mind
Michael Pollan Penguin Press *(2018)*
Journalist Michael Pollan explored psychoactive plants in *The Botany of Desire* (2001). In this bold, intriguing study, he delves further, homing in on psychedelic compounds currently under study, such as psilocybin. He meets a vast cast of researchers, including mycologist Paul Stamets and neuroscientist Robin Carhart-Harris, who works on neural correlates of the psychedelic experience. Pollan even "shakes the snow globe" himself, chemically self-experimenting in the spirit of psychologist William James, who speculated about the wilder shores of consciousness more than a century ago.



Still Waters
Curt Stager W. W. Norton *(2018)*
"There is nothing like a lake to reflect and reveal the world." So declares ecologist Curt Stager, whose lyrical evocation of 'living waters' offers geological and biological revelations. He also probes our relationship to lakes as one body of water to another, examining mud cores from Walden Pond in Massachusetts (immortalized by nineteenth-century naturalist Henry David Thoreau), flesh-dissolving alkaline minerals in Tanzania's Lake Natron, and the endemic species crowding crystal-clear Lake Baikal in Siberia. **Barbara Kiser**

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Correspondence

Recruit fresh talent for coral reefs

Coral reefs are again in the spotlight, having suffered mass mortality over the past two years from global bleaching events. Before reef resilience runs out, researchers must move beyond lamenting corals' lost pristine state and develop pragmatic solutions. In our view, these are likely to stem from a more diverse set of stakeholders than have participated so far.

We must ensure that reefs can continue to provide well-being for millions of people in the future, despite widespread alterations in their biological state. Degraded reefs still have the potential to provide fisheries benefits, cultural value and other sources of revenue (such as tourism), although all of these are likely to be reduced.

With 2018 designated the International Year of the Reef, fresh perspectives and approaches are needed (S. A. Hewlett *et al. Harvard Bus. Rev.* **91**, 30; 2013). New recruits should come from a greater variety of sectors (such as development, health and governance) and from a wider set of disciplines (such as the social sciences — including psychology, economics, political science and geography) than today's conservationists. Young scientists and researchers from the global south will be particularly important contributors.

Gabby Ahmadia* *World Wildlife Fund, Washington DC, USA.*
gabby.ahmadia@wwfus.org
**On behalf of 14 co-signatories (see go.nature.com/2kwzqk7 for full list).*

Atacama imperilled by lithium mining

The demand for lithium — used in rechargeable batteries for mobile phones, electric vehicles and other devices — caused a 13% surge in global production last year (go.nature.com/2guqzc8). In the Salar de Atacama in Chile, part of South America's vast 'lithium triangle' of high-altitude lakes and salt flats, more than 1,700 litres of lithium brines are pumped from the shallow subsurface every second. This intense activity in one of the driest areas in the world is causing serious friction over water rights between local communities and mining companies and is putting huge pressure on a fragile and poorly understood ecosystem.

For example, the region's isolated wetlands are rich in species that are unique to the area. These are crucial islands of habitat for migratory and resident birds, including the threatened Andean flamingo (*Phoenicoparrus andinus*). Harmful cyanobacteria usually eaten by these birds accumulate in the water polluted by lithium extraction, putting human health at risk (T. C. Wanger *Conserv. Lett.* **4**, 202–206; 2011).

A Chilean parliamentary commission has acknowledged that overexploitation of groundwater resources has damaged ecosystems in the Salar de Atacama basin, and that little attention has been paid to threats from mining (see go.nature.com/2mnhuwm; in Spanish). We urge the government to rethink its policies to account for the political, social and ecological impact of huge mining projects.

Jorge S. Gutiérrez, Juan G. Navedo *Austral University of Chile, Valdivia, Chile.*
Andrea Soriano-Redondo *University of Exeter, Penryn, UK.*
jorgesgutierrez@unex.es

Risks from extracted technology metals

The global market for rare metals from Earth's crust is on the rise because of their crucial role in technologies used in electronics, biomedicine and the automotive industry, for example. Most of these elements are extracted in African and South American countries, where environmental protections are often poor. We call for rigorous investigation into these metals' environmental concentrations, biogeochemical cycles and possible risks to human and animal health.

Technology-critical elements include tantalum, gallium, germanium, indium, niobium, tellurium, thallium and other scarce metals. Despite their economic importance (see, for example, A. L. Gulley *et al. Proc. Natl Acad. Sci. USA* **115**, 4111–4115; 2018), little is known about their environmental effects (M. Filella *Earth Sci. Rev.* **173**, 122–140; 2017). For example, tantalum increases in aquatic organisms at each successive level of the food chain (W. Espejo *et al. Environ. Sci. Technol. Lett.* **5**, 196–201; 2018) but it is unclear whether this accumulation poses a threat to humans and other consumers.

A better understanding of the effects of extraction and consumption of technology-critical elements will help to mitigate the risks to environmental and human health in producer countries.

Winfred Espejo *University of Concepción, Chile.*
Cristóbal Galbán-Malagón *Andres Bello University, Santiago, Chile.*
Gustavo Chiang Merimoyu *Ecosystem Research Institute Foundation (MERI), Santiago, Chile.*
cristobal.galban@unab.cl

Publish translations of Chinese papers

Language is still a barrier to scientific development (see, for example, V. S. Lazarev and S. A. Nazarovets *Nature* **556**, 174; 2018). We suggest that the best research papers published in Chinese or other languages (for instance, highly cited articles) should be routinely translated and republished to render them more visible to the English-language-dominated research community.

Since 1979, around 79 million papers have been published in Chinese — including in China's highest-quality journals, according to the China National Knowledge Infrastructure databases (http://oversea.cnki.net; see also *Nature* **553**, 390; 2018). Many important advances are therefore going unseen by Western researchers.

An example is a landmark study by Youyou Tu, who shared a Nobel prize in 2015 for the discovery of artemisinin and the treatment of malaria (Y. Tu *et al. Acta Pharm. Sin.* **16**, 366–370; 1981), which was cited only once outside China. And as of 2 May, all but 3 of 347 citations of the most-cited Chinese-language paper in the Web of Science Core Collection came from Chinese authors. (The paper discusses a radioisotope technique that is used to date rocks; see F. Y. Wu *et al. Acta Petrol. Sin.* **23**, 185–220; 2007).

Breakthroughs such as Microsoft's algorithm for Chinese–English machine translation could speed up international sharing of Chinese publications (see go.nature.com/2jhxuwo). Efforts need to focus on which papers should be selected for translation by engaging with publishers, authors and other experts, and on resolving copyright-ownership issues.

Juan Tao, Chengzhi Ding *Yunnan University, Kunming, China.*
Yuh-Shan Ho *Asia University, Wufeng, Taichung, Taiwan.*
chzhding@ynu.edu.cn

CONTRIBUTIONS

Correspondence may be submitted to **correspondence@nature.com** after consulting the author guidelines and section policies at **http://go.nature.com/cmchno**.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ASTROPHYSICS

Pulsars seen through a new lens

Radio waves produced by cosmic lighthouses called pulsars are distorted by surrounding material. Observations show that this material can act as a lens, focusing the waves and boosting the brightness of the pulsar. SEE LETTER P.522

JASON HESSELS

Much of what we know about the Universe comes from using telescopes to collect and study the electromagnetic waves produced by astronomical objects. These waves can be distorted and absorbed as they travel to Earth, providing valuable information about the otherwise invisible material that lies between stars and between galaxies. Rotating stellar remnants called pulsars have long been known to be excellent probes of interstellar matter[1]. A pulsar generates a beam of radio waves that flashes across the sky as the object rotates, just like the light from a lighthouse. On page 522, Main *et al.*[2] report that ionized gas close to a pulsar can greatly amplify the object's observed brightness, and allow astronomers to zoom in on the source of the emission. The discovery could also help to explain enigmatic astrophysical signals known as fast radio bursts.

Main and colleagues targeted a pulsar[3] that is officially termed PSR B1957+20, but is known informally as the black widow (Fig. 1). The pulsar and a companion star exist in a binary system, in which the two bodies orbit around a common centre of mass. The binary system itself is minuscule — it would almost fit inside the Sun.

In addition to its radio beam, the pulsar generates a strong wind of particles that strips matter from the companion star and produces a surrounding cloud of ionized and magnetized gas called a plasma. This is where the ominous name of the pulsar comes from — like the sexually cannibalistic spider after which it is named, the black-widow pulsar is destroying its mate. From the point of view of an observer on Earth, the pulsar is eclipsed once per orbit, when it travels behind the plasma cloud.

Main *et al.* used the 305-metre-diameter William E. Gordon Telescope at the Arecibo Observatory in Puerto Rico to study the black-widow pulsar in unprecedented detail. The authors meticulously mapped how the observed brightness of the pulsar changes on timescales as short as microseconds. What they found is that the pulsar seems to be much brighter at certain points of its orbit — in particular, at the edges of the plasma cloud, just before and after the pulsar is eclipsed. For only a few milliseconds at a time, and at specific radio frequencies, the pulses can appear to be up to 80 times brighter than average (see Fig. 2 of the paper[2]).

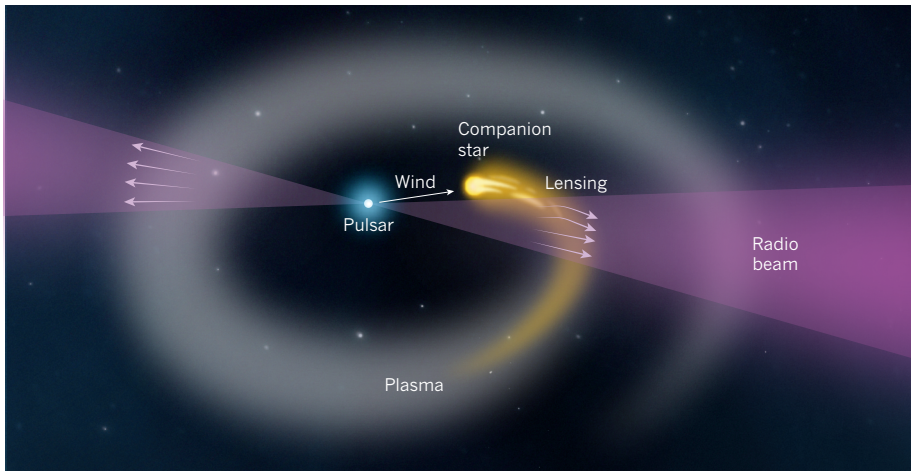


Figure 1 | The black-widow pulsar. A pulsar is a rotating stellar remnant that generates a powerful beam of radio waves. Main *et al.*[2] report observations of a binary system that comprises a pulsar known as the black widow and a small companion star. A wind of particles from the pulsar strips matter from the companion star and produces a surrounding cloud of ionized and magnetized gas called a plasma. The authors demonstrate that the plasma relatively close to the companion star can act as a lens, focusing the radio waves and boosting the observed brightness of the pulsar.

The authors hypothesize that the edges of the plasma cloud act as a lens, boosting the observed brightness of the pulsar for short periods. Think of how a magnifying glass focuses light; in a broader sense, any medium that causes electromagnetic waves to change direction can serve as a lens. In a plasma lens, radio waves are bent, and the waves arriving at an observer from different angles can overlap to produce a bright spot known as a caustic[4]. A similar effect with light can be seen at the bottom of a swimming pool on a sunny day.

Main and colleagues also demonstrated that the plasma-lensing effect can be used to zoom in on the pulsar, by studying how the effect changes with time and with observed radio frequency. This is astounding because pulsars are no more than about 30 kilometres in diameter[5]. They produce their radio waves from a magnetized atmosphere known as a magnetosphere, which rotates along with the pulsar. In the case of a fast-spinning pulsar such as the black widow, the magnetosphere extends only about 100 km above the pulsar's surface[3].

Based on the time- and frequency-dependent way in which the observed pulses are amplified, Main *et al.* inferred that they could detect the physical offset in emissions that originate from different regions of the magnetosphere. In other words, the lensing effect allowed the authors to measure a kilometre-scale offset despite the fact that the black-widow pulsar is about 10^{17} km away. Such a feat is equivalent to measuring the width of a hair on Mars from Earth.

The black-widow pulsar provides a stunning illustration of plasma lensing in action, but there are other astrophysical examples of this phenomenon. For instance, variations in the observed radio brightness of distant supermassive black holes called quasars have been ascribed to lensing from poorly understood plasma structures in our Galaxy[6,7]. Furthermore, echoes of pulses from the Crab pulsar could be caused by lensing from plasma filaments in the pulsar's surrounding nebula[8]. Again, plasma lensing not only boosts the brightness of astronomical objects, but also reveals what would otherwise be invisible.

Plasma lensing might allow astronomers to peer deeper into the Universe than would otherwise be possible. More specifically, Main and colleagues suggest that their observations provide a clue to solving the mystery of fast radio bursts — millisecond-duration radio flashes that seem to originate in distant galaxies, but

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

which are bright enough to detect on Earth.

Intriguingly, the effects observed in the black-widow pulsar are similar to distortions seen in pulses from at least one source of fast radio bursts[9]. Plasma lensing might therefore be responsible for boosting the brightness of fast radio bursts, as has previously been hypothesized and modelled[10]. However, the story is not complete: the environment around a source of fast radio bursts is probably quite different from, and possibly even more extreme than, that of the black-widow pulsar. It perhaps has more in common with the environment around the young Crab pulsar or at the centre of our Galaxy[11].

Main *et al*. have detected plasma lensing in a pulsar that has been studied for more than 30 years, using a telescope that has been operating since the early 1960s. Why the sudden insight? As computing and data-recording power has grown, so has the ability to use venerable radio telescopes to scrutinize pulsars on shorter timescales and over a wider range of radio frequencies. This suggests that the future is bright for using pulsars to illuminate the invisible Universe. ■

Jason Hessels *is at ASTRON (the Netherlands Institute for Radio Astronomy) and the Anton Pannekoek Institute for Astronomy, University of Amsterdam, 1098XH Amsterdam, the Netherlands.*

e-mail: j.w.t.hessels@uva.nl

1. Rickett, B. J. *Annu. Rev. Astron. Astrophys.* **28**, 561–605 (1990).
2. Main, R. *et al. Nature* **557**, 522–525 (2018).
3. Fruchter, A. S., Stinebring, D. R. & Taylor, J. H. *Nature* **333**, 237–239 (1988).
4. Clegg, A. W., Fey, A. L. & Lazio, T. J. W. *Astrophys. J.* **496**, 253–266 (1998).
5. Hessels, J. W. T. *et al. Science* **311**, 1901–1904 (2006).
6. Romani, R. W., Blandford, R. D. & Cordes, J. M. *Nature* **328**, 324–326 (1987).
7. Bannister, K. W. *et al. Science* **351**, 354–356 (2016).
8. Backer, D. C., Wong, T. & Valanju, J. *Astrophys. J.* **543**, 740–753 (2000).
9. Spitler, L. G. *et al. Nature* **531**, 202–205 (2016).
10. Cordes, J. M. *et al. Astrophys. J.* **842**, 35 (2017).
11. Michilli, D. *et al. Nature* **553**, 182–185 (2018).

DEVELOPMENTAL BIOLOGY

Rethinking WNT signalling

The identification of genetic mutations that can hinder the development of human limbs has led to the discovery of an unanticipated mode of regulation for the WNT signalling pathway during limb development. SEE LETTER P.564

JESSICA A. LEHOCZKY & CLIFFORD J. TABIN

After more than 30 years of intense study, the major intercellular signalling systems that orchestrate embryonic development and tissue maintenance are reasonably well understood. Although there are many crucial details yet to be worked out, there is a tendency among researchers in the field to think that the major players in these signalling pathways — and the ways in which they interact — are known. On page 564, Szenker-Ravi *et al.*[1] remind us that this is not necessarily the case.

WNT proteins are signalling molecules whose activity controls many processes, from tissue organization and body-axis formation during embryonic development to maintenance and regulation of stem cells in adult tissue. WNTs signal through several distinct intracellular pathways, but these pathways share an initial step: WNT molecules outside the cell bind to and activate receptors of the Frizzled family that span the cell membrane. Frizzled receptors must be present in the membrane for intracellular WNT-pathway activation.

In vertebrates, an auxiliary regulatory process is involved in controlling the accumulation of Frizzled receptors, and hence in determining WNT signalling levels. The system involves three groups of proteins[2,3]: an LGR (LGR4, 5 or 6), an extracellular R-spondin (RSPO1, 2, 3 or 4), and an E3 ubiquitin ligase enzyme (ZNRF3 or RNF43). In the absence of RSPOs, the ubiquitin ligase tags Frizzled receptors with ubiquitin molecules, which mark the receptors for degradation. This results in low membrane concentrations of Frizzled, so WNT signalling is attenuated[4–6] (Fig. 1a). Conversely, when RSPOs are present, they bind to LGRs[2,3], and the resulting complex binds the ubiquitin ligase and prevents it from tagging Frizzled. The receptors accumulate, and WNT signalling can occur[4–6] (Fig. 1b).

This mechanism has a crucial role in regulating WNT signalling in stem-cell compartments characterized by LGR expression, for instance in the intestine[7] and hair follicles[8]. Because WNT signalling has been extensively studied in these contexts, the LGR–RSPO–ligase complex has become part of the standard picture of how WNT activity is regulated.

Szenker-Ravi and colleagues' study began with a genetic analysis of five families affected by either tetra-amelia syndrome, which is characterized by the lack of all four limbs and by lung abnormalities, or by a previously undescribed syndrome involving severe limb malformations. The authors found that these two syndromes are caused by five mutations in the *RSPO2* gene that disrupt different protein domains. Using *in vitro* assays, the researchers demonstrated that the mutations prevent RSPO2 from binding to LGR or RNF43, and so inhibit WNT signalling.

So far, Szenker-Ravi and co-workers' data

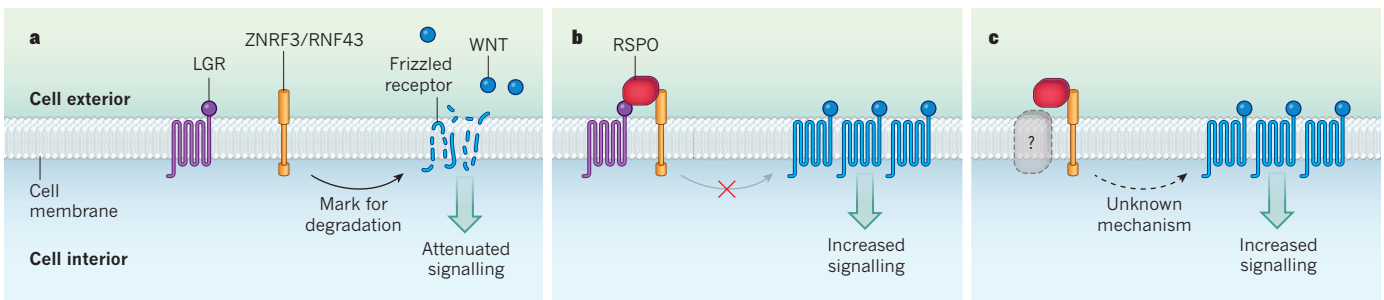


Figure 1 | Updated model of WNT-signalling regulation. WNT signalling, which is triggered when WNT proteins bind to Frizzled receptors that span the cell membrane, is conventionally thought to be regulated by interactions between three groups of proteins: R-spondins (RSPO1 to 4), LGRs (LGR4 to 6) and ubiquitin ligase enzymes (ZNRF3 or RNF43). **a**, In the absence of RSPOs, LGR and ZNRF3 or RNF43 do not interact, and the ubiquitin ligase catalyses a reaction that marks Frizzled receptors for degradation. WNTs cannot bind the degraded Frizzled, and WNT signalling is hampered. **b**, When present, RSPO binds LGRs and ubiquitin ligase, preventing enzyme activity. This allows Frizzled receptors to accumulate and so increases WNT signalling. **c**, Szenker-Ravi *et al.*[1] report that, in the developing limb and lung, RSPO can bind ubiquitin ligase without LGRs. They propose that another, unidentified, protein enables this interaction, which leads to increased WNT signalling through an unknown mechanism.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

fit nicely with the known roles of RSPOs, LGRs and ubiquitin ligases. And, like people carrying *RSPO2* mutations, mice lacking *Rspo2* have limb abnormalities[9]. The authors expected that the loss of LGR activity would have the same effect. But they got a surprise when they analysed mice lacking the *Lgr4*, *5* and *6* genes — the triple-mutant embryos did not have limb or lung abnormalities. This suggests that, in some tissues, RSPO2 (and perhaps other RSPOs) can act independently of LGRs, potentiating WNT signalling in the absence of its usual binding partner.

To test this idea directly, the group next investigated whether cells isolated from LGR triple-mutant embryos are capable of RSPO-mediated WNT signalling. They found no evidence of WNT signalling when these cells were exposed to *Rspo1* or *Rspo4*, but WNT activity was detected in the presence of *Rspo2* or *Rspo3*. Thus, RSPO2 and RSPO3 seem to be able to induce WNT signalling independently of LGRs. However, these RSPOs still seem to act through their normal ubiquitin ligase targets, because Szenker-Ravi *et al.* found that modulation of ZNRF3 alters WNT signalling in triple-mutant cells. Consistent with this picture, the authors showed that deletion of *rspo2* in the frog *Xenopus laevis* led to missing limbs, whereas deletion of the *znrf3* and *rnf43* genes led to extra limbs.

This study demonstrates that the accepted model of WNT-receptor modulation does not hold in the case of limb and lung development. Szenker-Ravi *et al.* hypothesize that a separate, unidentified receptor is necessary for this LGR-independent WNT signalling (Fig. 1c). Notably, a study published earlier this year[10] identified one potential candidate. That work showed that RSPO2 and RSPO3 can bind to ZNRF3 or RNF43 in conjunction with heparin sulfate proteoglycan (HSPG) molecules in lieu of LGRs, to enable WNT signalling *in vitro*. Future work will be required to test whether HSPGs play this part in the context of lung and limb development. In addition, it remains to be determined whether the HSPG–RSPO–ZNRF3 complex promotes WNT signalling by preventing ZNRF3 activity, or whether another mechanism is at work. Either way, it will be important to determine the extent of any functional similarities between LGR- and HSPG-based complexes, and to uncover whether there is any pattern to the use of LGR or HSPG as a cofactor in a particular tissue.

Szenker-Ravi and colleagues' work also points to ways to broaden our understanding of processes that require WNT signalling, such as limb development. For example, analysis of the early stages of limb development in frog embryos lacking *znrf3* and *rnf43* could reveal why these mutations lead to extra limbs. Do ZNRF3 and RNF43 act as 'master regulators' of limb numbers, as the authors propose? Consistent with this idea, WNT activity has a role in initiating the formation of the limb bud[11] (which eventually gives rise to the limb). Alternatively, rather than being master regulators, these proteins might mediate limb numbers indirectly. For example, extra limbs might arise as a secondary consequence of expansion of the pool of limb progenitor cells, or they might arise because of changes in the formation of a signalling centre at the tip of the limb bud that directs limb outgrowth — both WNT-dependent processes[12,13].

Finally, it will be interesting to evaluate LGR-independent, RSPO-mediated WNT signalling in cancer. Chromosomal abnormalities that lead to activation of RSPO2 or RSPO3 have been shown to drive WNT-dependent colon tumours[14]. Szenker-Ravi and colleagues' demonstration that these two RSPOs can modulate WNT activity independent of LGR adds a twist to these findings, and should prompt scientists to look for cancer-causing mutations in *RSPO2* or *RSPO3* in cells outside LGR-expressing cell compartments. ■

Jessica A. Lehoczky *is in the Department of Orthopedic Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA.* **Clifford J. Tabin** *is in the Department of Genetics, Harvard Medical School.*
e-mails: jlehoczky@bwh.harvard.edu; tabin@genetics.med.harvard.edu

1. Szenker-Ravi, E. *et al. Nature* **557**, 564–569 (2018).
2. Chen, P.-H., Chen, X., Lin, Z., Fang, D. & He, X. *Genes Dev.* **27**, 1345–1350 (2013).
3. Gong, X. *et al. PLoS ONE* **7**, e37137 (2012).
4. Hao, H.-X. *et al. Nature* **485**, 195–200 (2012).
5. Carmon, K. S., Gong, X., Lin, Q., Thomas, A. & Liu, Q. *Proc. Natl Acad. Sci. USA* **108**, 11452–11457 (2011).
6. de Lau, W. *et al. Nature* **476**, 293–297 (2011).
7. Sato, T. *et al. Nature* **459**, 262–265 (2009).
8. Jaks, V. *et al. Nature Genet.* **40**, 1291–1299 (2008).
9. Bell, S. M. *et al. Development* **135**, 1049–1058 (2008).
10. Lebensohn, A. M. & Rohatgi, R. *eLife* **7**, e33126 (2018).
11. Kawakami, Y. *et al. Cell* **104**, 891–900 (2001).
12. ten Berge, D., Brugmann, S. A., Helms, J. A. & Nusse, R. *Development* **135**, 3247–3257 (2008).
13. Kengaku, M. *et al. Science* **280**, 1274–1277 (1998).
14. Han, T. *et al. Nature Commun.* **8**, 15945 (2017).

This article was published online on 16 May 2018.

ANTHROPOLOGY

Sizing up human brain evolution

An innovative computational analysis of factors that might have influenced human brain evolution suggests that ecological, rather than social, factors had a key role in the evolution of large, rapidly developing brains. SEE LETTER P.554.

RICHARD MCELREATH

Most organisms are brainless but thriving. Brains are expensive to produce and maintain, and in the human lineage they have grown so large as to incur a substantial metabolic burden as the brain develops[1]. A human brain stops growing by the age of ten, long before the body reaches physical maturity, and this costly and fast process of brain growth has been proposed to cause a delay in body growth[1]. Brain growth is not given priority in this way in other apes, and the human pattern is puzzling because it keeps our bodies smaller, more vulnerable and less productive for longer. The answer to this riddle must lie in how the human brain helped our ancestors to survive and reproduce. On page 554, González-Forero and Gardner[2] investigate the role of different factors as possible drivers of our unusually large brains, and determine how well these factors might account for the pattern of changes in brain and body size that occur as humans develop.

Proposals for how large brains evolved in humans include ecological, social and cultural hypotheses. The ecological-intelligence hypothesis suggests that environmental challenges, such as finding food, are paramount in driving brain-size evolution[3]. The social-intelligence hypothesis suggests instead that the competitive and cooperative challenges of living with other members of the same species are the key factor[4]. The cultural-intelligence hypothesis combines these two ideas, suggesting that the social learning of ecologically relevant skills explains the extreme brain investment of our lineage[5].

Until now, testing these hypotheses has relied mainly on comparative studies that correlate data on brain characteristics such as size (as an approximation of intelligence) with features such as cognition, ecology and group living. These regression approaches, which seek to identify variables that are associated with brain size, have been valuable for refining theories and the data measurements needed.

However, such regression studies can generate conflicting and confusing results. Changes to brain and body growth can have a reciprocal effect on each other for various reasons, such as metabolic constraints and energy-production needs, so such interactions between the brain and the body are complex and nonlinear. This makes the results of regression studies hard to interpret, because they cannot be connected directly to a relevant

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

evolutionary model. Researchers in the field should stop theorizing using one set of models while analysing data with another. Moving from purely statistical models, such as regression approaches, to studies that test evolutionary models could accelerate future progress.

The study of human brain evolution must by necessity be observational, because direct experimentation to test the role of variables is not an option. But working out what affects different components in such observational systems is hard. When Ronald Fisher, a leading evolutionary biologist and statistician of the twentieth century, was asked how one could infer causality in such cases, his advice was to "make your theories elaborate"[6].

Automobile engineering can provide an analogy for studying this type of system. It would be difficult to understand racing-car design through regression analysis of how engine size varies depending on changes in other features, such as the mass and shape of the car. Instead, a model is needed that uses physical laws to predict optimal combinations of the variables under different criteria. Understanding brain evolution poses a similar challenge in that an organism's features co-evolve under biological constraints.

González-Forero and Gardner's approach heeds Fisher's advice because the authors generated an elaborate model to investigate brain evolution. Modelling brain evolution in this way can produce many precise predictions of brain size that can easily be falsified. And because the model is based on biological characteristics, it is easy to learn from it. When the model's results do not match the observed evidence of brain size, the biological assumptions can be studied to understand why the model failed.

In the authors' computational set-up, as a human individual ages, there is a schedule of investment in brain, body and reproductive tissue. As individuals grow, an increase in brain size allows for an increase in skill, and an increase in body size makes it easier to convert that skill into energy. The skill boost also aids successful reproduction. The model generates life-history scenarios that are linked to specific predictions of brain and body sizes.

The metabolic costs of maintaining bodies and brains were assigned in the model by using previously determined metabolic-scaling relationships, which provide information such as how the metabolic rate changes depending on an organism's size. These metabolic costs were fixed in the authors' model, and the importance of different types of challenge were estimated by varying the weighting of these challenges and assessing the subsequent effect on the predicted brain and body sizes (Fig. 1). The authors explored four types of challenge: ecological (me versus nature), cooperative ecological (us versus nature), between-individual competitive (me versus you), and between-group competitive (us versus them). The authors determined which combination of challenge weighting gave rise to a pattern of hypothetical brain and body growth that was most consistent with that observed during human life history.

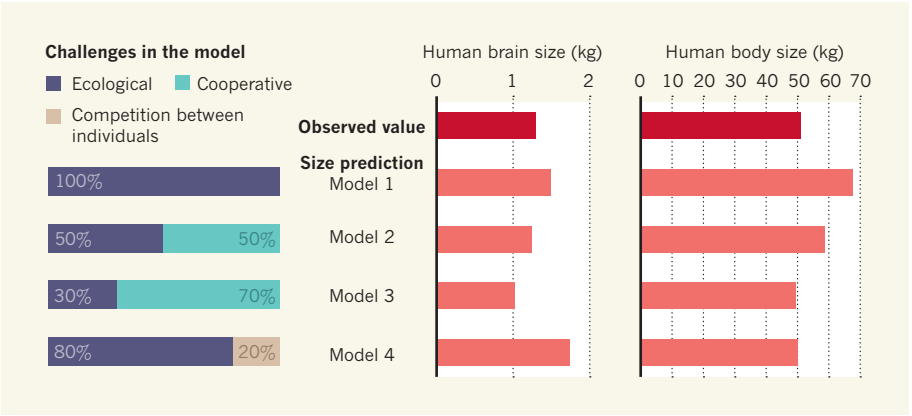


Figure 1 | Modelling the evolution of human brain size. Compared with other apes, humans have distinctively large and rapidly developing brains[1], and how this human developmental pattern evolved is debated. González-Forero and Gardner[2] report a computational modelling analysis that investigates the role of ecological factors and social factors (such as cooperation or competition between individuals) in driving the evolution of human brain size. The authors' model predicts human brain and body size depending on the relative weighting of ecological and social factors. Some examples of challenge weighting are shown to the left of the corresponding predictions generated in modelling results (data from Fig. 3 of ref. 2). Comparing such predicted values with the observed average brain and body size of a female adult enabled the authors to determine the relative importance of the evolutionary drivers, leading them to identify ecological drivers as being the major determinant of human brain size in their analysis.

González-Forero and Gardner's analysis reveals a major role for ecological intelligence in driving human brain and body growth in this system. The best match between model predictions and observed human growth patterns came from assigning a weight of 60% to ecological challenges in the model.

By contrast, social challenges were less likely to contribute to the observed human growth patterns. Competitive challenges between individuals or groups are linked to large brains and to a body size that is smaller than the observed value. In competition, as skill increases, such gains in skill can lead to diminishing returns in terms of an increase in energy extraction because what each individual is competing against becomes continually harder to overcome. For example, skill increases in one individual could be matched by skill increases in other competitors, thereby limiting the energy boost from a skill increase. By contrast, the challenge itself doesn't evolve in ecological challenges, so overcoming ecological challenges can lead to a more-efficient energy boost. The best-matched model had a 10% weight for between-group competition.

Cooperation was found to have more of an effect. The best-matched model assigns 30% weight to cooperative challenges. However, cooperation could lead to a reduction in brain size because individuals could potentially free-load on the intelligence of others, evidence of this effect has been observed in some animals[7].

Ecological drivers are the clear winner. But the model fails to address the possible role of cultural intelligence, as the authors admit, because cultural dynamics are not included. The authors' results are consistent with the cultural-intelligence hypothesis, but any such possible connection remains speculative.

Some of the model's size-prediction results are sensitive to the details, such as the precise way in which skill translates into reproductive success. However, this provides a valuable opportunity to understand previously unnoticed implications of hypotheses about the challenges driving brain evolution, and to identify targets for future work. For example, the model would benefit from more measurements of the rate at which skills increase with age, because few data of this kind currently exist.

Finally, because the model aims to explain brain size in humans only, the results have no clear significance for debates about the evolution of intelligence in other species. Nevertheless, the methodological implications of this work are enormous. This type of general framework to investigate and predict values for constellations of co-evolving variables, not only in adults but also throughout life, would allow for more-detailed tests of more-nuanced predictions, regardless of the species of interest. ■

Richard McElreath *is in the Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany.*
e-mail: richard_mcelreath@eva.mpg.de

1. Kuzawa, C. W. *et al. Proc. Natl Acad. Sci. USA* **111**, 13010–13015 (2014).
2. González-Forero, M. & Gardner, A. *Nature* **557**, 554–557 (2018).
3. Parker, S. T. & Gibson, K. R. *J. Hum. Evol.* **6**, 623–641 (1977).
4. Humphrey, N. K. in *Growing Points in Ethology* (eds Bateson, P. P. G. & Hinde, R. A.) 303–317 (Cambridge Univ. Press, 1976).
5. Herrmann, E., Call, J., Hernàndez-Lloreda, M. V., Hare, B. & Tomasello, M. *Science* **317**, 1360–1366 (2007).
6. Cochran, W. G. *J. R. Stat. Soc. A* **128**, 134–155 (1965).
7. Thornton, A. & McAuliffe, K. *J. Zool.* **295**, 12–22 (2015).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

FORUM ECONOMICS

The cost of a warming climate

A study finds that meeting climate-change mitigation targets will lead to a substantial reduction in economic damages. Here, economists present opposing views on the approach used by studies such as this one. SEE LETTER P.549

THE TOPIC IN BRIEF

- Climate change is already affecting the economy through hurricanes, droughts and floods.
- On page 549, Burke *et al.*[1] report that achieving global-warming targets set by the United Nations could save trillions of dollars in damages.
- The study's methodology follows previous literature[2,3] by examining the short-term effects of weather on the growth rate of gross domestic product (GDP) — the market value of all goods and services produced in a country in a specific time period.
- These data are then extrapolated into the future to assess the economic impacts of climate change.
- The validity of this approach has been intensely debated in the economics community.

Feeling the heat

WOLFRAM SCHLENKER

Estimates of the economic impacts of climate change are essential for the development of climate policies. Important concerns have been raised about studies such as that of Burke *et al.*, and more research needs to be carried out. However, I think that the authors of these studies are doing the best job possible by basing their estimates on a rigorous analysis and clearly stating their assumptions.

The extrapolation of the historical relationship between temperature and GDP into the future raises the question of whether technological advances might change the predicted trajectory. However, it is worth emphasizing that such extrapolation is based on an econometric model — an economic model based on an empirical analysis — that has been shown to be remarkably consistent between rich and poor countries, as well as between the earlier and later part of the sample period involved (see, for example, ref. 3). This makes it unlikely that adaptation measures are already available, because they have not been deployed in the past even though hot countries would have benefited from them.

GDP is a useful metric to assess the benefits of limiting global warming. It provides a measurement of human welfare under the assumption that the market prices of goods and services fully reflect the costs of their production and use[4]. In reality, this assumption is not always valid. For instance, fossil-fuel prices often do not include the costs associated with global warming and other environmental effects on society. However, by focusing only on GDP, huge economic impacts from limiting global warming are predicted. These estimates would be even bigger if the non-market benefits of reduced fossil-fuel use — for example, for human health and ecosystems — were considered.

The predicted impacts are larger than those obtained in earlier work[5]. The main reason is that strong effects of weather on the growth rate of GDP are found, whereas the earlier work stipulated, but did not empirically test, that weather affects only the level of GDP in a particular year. Heat and drought, for example, directly influence agricultural yields in a given year, but have limited impact in the following years. By contrast, a growth effect implies that a destructive weather event not only decreases GDP in a given year, but also lowers the value for all future years.

The main innovation of studies such as that of Burke and colleagues is to use an econometric model that can incorporate both level and growth effects, without favouring one type of effect over the other[2]. This is accomplished by including both the current temperature and the temperatures in previous periods in the analysis.

If weather affected GDP only in a given year, immediate impacts would be offset in the future — for example, a 1% decrease in GDP would be offset by a 1% increase in GDP the following year. On the contrary, the authors of these studies find that the impacts are not offset, but rather amplified. The one caveat is that when temperatures in previous periods are included in the analysis, the uncertainties in the projected economic damages increase substantially. Resolving this issue is a key direction for future research. ■

Wolfram Schlenker *is at the School of International and Public Affairs and the Earth Institute, Columbia University, New York, New York 10027, USA.*
e-mail: wolfram.schlenker@columbia.edu

Grand damage projections

MAXIMILIAN AUFFHAMMER

Translating the impacts of climate change on surface temperature, precipitation and sea level into economic damages is challenging. One approach is to use longitudinal data (repeated measurements of the same variables) to estimate damage functions — mathematical expressions that translate physical impacts into monetary damages[6]. These functions are associated with specific locations and sectors, such as agriculture or manufacturing. A major drawback is that coverage across locations and sectors is incomplete. Studies such as that of Burke *et al.* circumvent this problem by using GDP to aggregate economic impacts across sectors. Nevertheless, several issues must be considered before strong conclusions can be drawn from this work.

First, the authors of these studies argue that societal adaptation to climate change is accounted for statistically. However, what is incorporated stems from a historical cross-country comparison of the temperature sensitivity of GDP. In reality, future adaptation will probably involve innovative technologies with lower costs than those that are currently used. Such technologies might include, for example, air conditioners powered by carbon-free electricity that are more energy efficient than present-day devices. Adaptation could therefore result in lower economic damages than are predicted.

Second, on a larger scale, climate change will lead to a planetary redistribution of economic activity, which will result in a redistribution of international trade flows. Such an effect is impossible to quantify credibly and could have a large impact on the projected damages.

Third, GDP includes only goods and services that are transacted in markets and therefore have measurable prices. It does not capture the effects of climate change on valuable non-market sectors, such as ecosystem services and biodiversity.

Finally, allowing climate change to influence both the level and the growth rate of GDP is shown to lead to a wide distribution of projected impacts. As a result, neither small (or zero) effects nor massive effects can be ruled out. Attempts to distinguish between these two

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

possibilities using simple statistical models at the global level have been inconclusive.

There are good reasons for thinking that some effects of climate change might be cumulative. For instance, climate and weather will affect the level, and potentially the growth rate and efficiency, of capital and labour. Furthermore, climate might induce technological change through both adaptation and mitigation measures. Pinning down these macroeconomic processes to resolve just how large the effects of climate will be on the long-term growth of GDP needs to be a high priority for future work. ■

Maximilian Auffhammer *is in the Department of Agricultural & Resource Economics, University of California, Berkeley, and the National Bureau of Economic Research, Berkeley, California 94720, USA.*
e-mail: auffhammer@berkeley.edu

1. Burke, M., Davis, W. M. & Diffenbaugh, N. S. *Nature* **557**, 549–553 (2018).
2. Dell, M., Jones, B. F. & Olken, B. A. *Am. Econ. J. Macroecon.* **4**, 66–95 (2012).
3. Burke, M., Hsiang, S. M. & Miguel, E. *Nature* **527**, 235–239 (2015).
4. Weitzman, M. L. *Q. J. Econ.* **90**, 156–162 (1976).
5. Nordhaus, W. D. *Am. Econ. Rev.* **83**, 313–317 (1993).
6. Carleton, T. A. & Hsiang, S. M. *Science* **353**, aad9837 (2016).

NEURODEGENERATION

Sabotage by the brain's supporting cells

Several neurodegenerative disorders are linked to the build-up of abnormal α-synuclein protein in distinct cell types. It emerges that differing intracellular factors dictate the properties of this protein in each cell type. SEE LETTER P.558

LARY C. WALKER

One of the many mysteries surrounding neurodegenerative diseases is how they can manifest in such a variety of ways. Disorders such as Alzheimer's disease, Parkinson's disease and amyotrophic lateral sclerosis (also known as motor neuron disease) are each defined by a core set of nervous-system abnormalities, but every affected person's brain responds slightly differently. Moreover, although each of these disorders is associated with the abnormal accumulation of a different protein in or around cells, some protein aggregates can give rise to more than one neurodegenerative disease. How can this happen? On page 558, Peng *et al.*[1] present persuasive evidence that different types of cell accumulate structurally distinct forms of one protein, α-synuclein. By shaping the 3D architecture of the corrupted protein, the cell type helps to determine the nature of the resulting disease.

Most normal proteins fold into characteristic conformations that are strongly governed by the protein's amino-acid sequence. But in age-related neurodegenerative conditions, certain proteins misfold, and induce other proteins of the same type also to misfold and to stick to one another. In this way, the abnormal molecular structure propagates by means of a crystallization-like process called seeded protein aggregation[2].

One such protein is α-synuclein. Under normal circumstances, α-synuclein is located mainly in nerve terminals. But in some cases, the protein forms intraneuronal aggregates called Lewy bodies and Lewy neurites — for instance in Parkinson's disease and a condition known as dementia with Lewy bodies, which are collectively referred to as Lewy body diseases (LBDs). In a more-aggressive brain disorder called multiple system atrophy (MSA), misfolded α-synuclein accumulates mostly in neuron-supporting cells called oligodendrocytes[3], in clumps known as glial cytoplasmic inclusions.

Why α-synuclein aggregates are mainly found in different cell types in MSA and LBDs has been uncertain. It cannot be attributed to differences in amino-acid sequence, because α-synuclein is not typically mutated in the common form of either condition[4,5]. However, previous work[6] has shown that aberrant α-synuclein in LBDs is structurally and functionally different from that in MSA. These variant molecular states are known as protein strains[7]. When injected into the brains of susceptible mice, the MSA strain causes a fatal disease similar to human MSA. By contrast, injecting the LBD strain fails to induce major signs of disease in this model[8].

Peng *et al.* set out to investigate the causes behind this difference in α-synuclein potency. The authors first confirmed that protein aggregates in the oligodendrocytes of people with MSA are conformationally distinct from those in neurons from people who have LBDs. In MSA, a few neurons do harbour α-synuclein aggregates, but the researchers found that these aggregates display the LBD conformation — thus, the two strains can occupy the same brain, albeit in different cell types. Next, the team exposed cultured cells to each strain, and found that MSA-derived α-synuclein is approximately 1,000 times more potent at inducing aggregation than is the LBD-derived protein.

The authors then injected the two types of aggregated α-synuclein (called seeds) into the brains of wild-type mice. This *in vivo* experiment confirmed that MSA-derived seeds are much more effective than seeds derived from LBDs at seeding aggregation. However, the seeds instigated aggregation only in neurons, not in oligodendrocytes.

Why might this be the case? Oligodendrocytes normally produce little, if any, α-synuclein[9]. The authors therefore genetically engineered mice to express α-synuclein only in oligodendrocytes. They found that α-synuclein aggregation could be induced in oligodendrocytes in these mice using seeds from either the MSA or the LBD strain — but again, the MSA strain was much the more potent. Importantly, the aggregates that emerged were always the MSA strain, regardless of the type of seed injected. Finally, when Peng *et al.* exposed synthetic, unaggregated

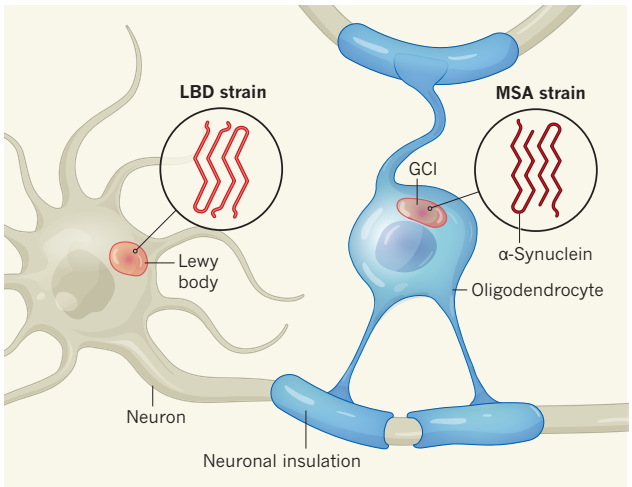


Figure 1 | Distinct strains of α-synuclein protein. In Parkinson's disease and dementia with Lewy bodies (collectively referred to as Lewy body diseases; LBDs), a misfolded form of α-synuclein called the LBD strain aggregates mainly in neurons to form anomalous structures called Lewy bodies and Lewy neurites (not shown). But in a disease known as multiple system atrophy (MSA), a different strain of misfolded α-synuclein forms aggregates called glial cytoplasmic inclusions (GCIs) in oligodendrocytes — non-neuronal cells that normally produce the fatty insulation for neuronal projections. Peng *et al.*[1] show that differences in the intracellular environments of the two cell types are responsible for the formation of the two strains.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

α-synuclein to cellular material that had been isolated from ruptured oligodendrocytes or neurons, the oligodendrocytic homogenate caused the protein to aggregate into a strain of greater seeding potency than did the neuronal homogenate. Some as yet unidentified factors in oligodendrocytes, then, seem to drive the formation of the MSA strain.

Collectively, the authors' experiments show that oligodendrocytes produce a structural variant of α-synuclein that robustly seeds aggregation in both oligodendrocytes and neurons. However, the cell type strongly influences the molecular strain of the aggregates that forms: oligodendrocytes specifically generate the MSA strain, whereas neurons preferentially produce the LBD strain (Fig. 1). The authors conclude that cell-type-specific factors regulate the molecular architecture and distinctive harmful properties of aberrant α-synuclein aggregates. They further suggest that the robust seeding capacity of the MSA strain contributes to the aggressive clinical progression of MSA.

The recognition that different cells generate different disease-causing protein strains raises a wealth of questions. An especially perplexing problem is how such large amounts of α-synuclein end up in oligodendrocytes in MSA, given that these cells do not produce much of the protein[9]. Possible mechanisms include augmented oligodendrocytic expression of α-synuclein in the disease state, or the uptake of α-synuclein that has been released from neurons[10].

MSA can affect different brain systems in people with the condition[3]. Peng and colleagues' findings demonstrate the possible role of cell composition in these differences, but what other factors might be involved? Perhaps the metabolism of α-synuclein in neurons and oligodendrocytes varies between brain regions. Or maybe the regional vulnerability to α-synuclein deposition is dictated mainly by where seeding first occurs. Two other types of neuron-supporting cell, astrocytes and microglia, are also involved in MSA, but their roles in the disease remain to be defined[11]. Finally, the role of small, toxic α-synuclein assemblages called oligomers[12], which might act in different brain regions from the larger aggregates, needs further exploration. Clarification of these issues could improve our understanding not just of α-synuclein diseases, but also of other neurodegenerative disorders that involve protein aggregation.

In sum, the authors' work highlights the interplay between a cell's composition and proteins that can cause disease. Identification of the cell-specific factors that impel α-synuclein to aggregate into the MSA strain could reveal ways to treat this debilitating and ultimately fatal brain disorder. Peng and colleagues' findings are also a salutary reminder of the salient part played by non-neuronal cells in both the health and failure of the nervous system. ■

Lary C. Walker *is in the Department of Neurology and Yerkes National Primate Research Center, Emory University, Atlanta, Georgia 30322, USA.*
e-mail: lary.walker@emory.edu

1. Peng, C. *et al. Nature* **557**, 558–563 (2018).
2. Jucker, M. & Walker, L. C. *Nature* **501**, 45–51 (2013).
3. Krismer, F. & Wenning, G. K. *Nature Rev. Neurol.* **13**, 232–243 (2017).
4. Nussbaum, R. L. *Cold Spring Harb. Perspect. Med.* **7**, a024109 (2017).
5. Hernandez, D. G., Reed, X. & Singleton, A. B. *J. Neurochem.* **139** (Suppl. 1), 59–74 (2016).
6. Peelaerts, W. *et al. Nature* **522**, 340–344 (2015).
7. Walker, L. C. *Annu. Rev. Genet.* **50**, 329–346 (2016).
8. Prusiner, S. B. *et al. Proc. Natl Acad. Sci. USA* **112**, E5308–E5317 (2015).
9. Koga, S. & Dickson, D. W. *J. Neurol. Neurosurg. Psychiatry* **89**, 175–184 (2018).
10. Woerman, A. L. *et al. Cold Spring Harb. Perspect. Med.* **7**, a024588 (2017).
11. Brück, D., Wenning, G. K., Stefanova, N. & Fellner, L. *Neurobiol. Dis.* **85**, 262–274 (2016).
12. Bengoa-Vergniory, N., Roberts, R. F., Wade-Martins, R. & Alegre-Abarrategui, J. *Acta Neuropathol.* **134**, 819–838 (2017).

This article was published online on 9 May 2018.

CANCER

Subtype switch foils pancreatic tumours

Mutations in the gene *KDM6A* drive an aggressive subtype of pancreatic cancer by causing repositioning of an enzyme complex that modifies histone proteins associated with DNA, leading to altered gene expression.

FIEKE FROELING & DAVID TUVESON

Pancreatic ductal adenocarcinoma (PDAC) is one of the most deadly cancers in Western society[1] because it tends to be diagnosed late and responds poorly to therapy. PDAC tumours fall into two main groups[2–4]: a 'classical' subtype, and a more aggressive 'squamous' subtype in which pancreatic cells fail to undergo proper differentiation. The squamous subtype often involves mutations in members of the COMPASS-like complex — a group of methyltransferase and demethylase enzymes that, respectively, add or remove methyl groups from lysine amino-acid residues on histone proteins, around which DNA is packaged. Such histone modification can lead to changes in the expression of histone-associated genes involved in pancreatic-cell differentiation. Writing in *Cancer Cell*, Andricovich *et al.*[5] demonstrate the role of mutations in one member of this complex, KDM6A, in driving the squamous PDAC subtype.

The *KDM6A* gene is found on the X chromosome, and, in males, the presence of a *KDM6A* mutation can co-occur with mutation of a related gene on the Y chromosome, *UTY*. Andricovich *et al.* found that *KDM6A* and *UTY* mutations were associated with the squamous subtype of PDAC, and with shortened length of patient survival. They then used mouse models to confirm that functional KDM6A acts to suppress the development of PDAC. Mice harbouring *Kdm6a* gene mutations developed aggressive, poorly differentiated squamous tumours that showed protein- and gene-expression patterns characteristic of human tumours of the squamous subtype[2]. These defects were more pronounced in females than in males, consistent with the fact that females carry two copies of *Kdm6a* in their cells and males have one copy of *Kdm6a* on the X chromosome and *Uty* on the Y.

KDM6A is a demethylase that removes a methyl modification dubbed H3K27me3 from lysine residue 27 on histone H3 proteins. But Andricovich and colleagues found that only a small percentage of H3K27me3 marks were altered in cells from *Kdm6a*-mutant mice, compared with controls. This observation led the authors to posit that altered KDM6A demethylase activity was not the driver of *Kdm6a*-mutant PDAC. Instead, they found that loss of *Kdm6a* resulted in changes in other histone modifications, specifically at groups of gene-regulatory DNA sequences called super-enhancers, whose activation promotes the expression of certain genes that are highly expressed in PDAC.

In particular, the researchers observed changes in the distribution across super-enhancers of a different methyl modification (dubbed H3K4me1) and a modification involving acetyl groups at lysine 27 of histone H3 (H3K27ac). Such changes were associated with a repositioning of the COMPASS-like complex to these regions. The authors found that the altered histone modifications led to activation of some super-enhancers, and, in some cases, to an increased reach — an ability to regulate distant genes that the super-enhancer cannot normally influence. These findings indicate that KDM6A exerts its tumour-suppressive role not only through its demethylase activity, but also by altering

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

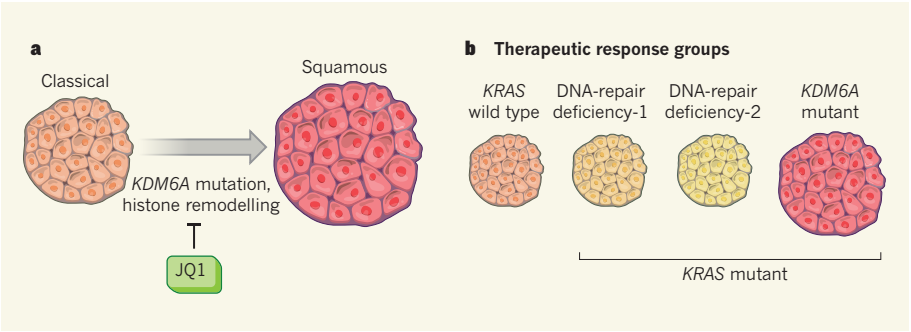


Figure 1 | KDM6A protein in pancreatic cancer. **a**, Pancreatic ductal adenocarcinomas (PDACs) can be catagorized into two main subtypes: classical tumours and more-aggressive, squamous tumours. Andricovich *et al.*[5] have provided evidence from mice and humans that mutations in the gene *KDM6A* cause changes in the patterns of molecular modifications to histone proteins, around which DNA is packaged. This histone remodelling leads to the expression of genes associated with squamous PDAC. However, the authors show that treatment with a small molecule called JQ1 prevents this subtype switch. **b**, This finding adds to the list of PDAC subgroups that can be targeted with drug treatments. Most PDAC tumours involve mutations in the gene *KRAS*, which cannot be targeted by drugs, but some *KRAS* wild-type tumours lack this mutation and carry other mutations that can be targeted. In addition, two subgroups of *KRAS*-mutant tumours carry defects in DNA-repair pathways, which can be targeted by different drugs.

the position of the COMPASS-like complex, enabling other enzymes to modify histones. The authors also showed that the increase in the reach and activation of super-enhancers led to the activation of genes involved in squamous-subtype-like differentiation.

Because *Kdm6a*-mutant PDAC in mice was not associated with significant H3K27me3 demethylation, the authors hypothesized that the alternative functions of the aberrant COMPASS-like complex promoted PDAC, and might therefore be vulnerable to drug treatment. This hypothesis is supported by the fact that mutant UTY, which helps to drive PDAC in males, lacks demethylase activity. Andricovich *et al.* therefore analysed the ability of various drugs that target other histone modifications to prevent the growth of *KDM6A*-deficient human cancer cells *in vitro*.

The authors found that cells harbouring mutations in *KDM6A* or other genes of the COMPASS-like complex were highly sensitive to inhibitors of BET-family proteins. These proteins bind to histone lysine residues that have been modified by acetyl groups, and recruit the cell's transcriptional machinery to promote gene expression. Various studies[6] have shown that BET inhibitors can displace the BET protein BRD4 from acetylated lysines at super-enhancer regions, thereby reducing the expression of cancer-causing genes (oncogenes) such as *MYC*. Because *KDM6A* mutations lead to altered lysine acetylation at super-enhancers, it makes sense that these drugs could be effective in this setting. Indeed, Andricovich *et al.* showed that the BET inhibitor JQ1 decreased BRD4 binding to the super-enhancers that regulate *MYC* and other oncogenes, and so decreased the expression of these genes.

Finally, the authors demonstrated that this drug treatment was also effective *in vivo*. The tumours of *Kdm6a*-deficient mice treated with JQ1 were smaller than those of mice that did not receive the drug, and had well-differentiated features typical of the classical PDAC subtype. This indicates that BET inhibitors have the potential to reverse the effects of the histone-modification remodelling that occurs in the squamous subtype (Fig. 1a). Targeting histone modifications and altered gene-regulatory networks to cause a 'class switch' to a more differentiated, less aggressive subtype of cancer might provide a promising therapeutic strategy. In support of the idea that modulating these factors can alter cancer progression, other studies have shown that enhancer reprogramming and large-scale losses of DNA methylation play a part in the spread of cancer[7,8].

Our increased understanding of the molecular underpinnings of cancer has hugely improved treatments for many tumours, although in PDAC the relative lack of obvious drug targets has presented a challenge. There are some cases of PDAC that involve oncogenes for which inhibitors do exist[9]. However, most PDAC tumours harbour oncogenic mutations in the gene *KRAS*, for which inhibitors are not available. But there are two clear groups of people with *KRAS*-mutant PDAC tumours characterized by deficiencies in specific DNA-repair pathways that can be targeted by drugs[10,11] (Fig. 1b). Patients harbouring *KDM6A* mutations (and possibly other mutations in genes of the COMPASS-like complex) might represent another subgroup, who would benefit from therapies targeting BET function. Moreover, BET inhibitors could have broader activity if combined with other inhibitors of histone remodelling, as previously reported[12].

It is to be hoped that more molecular biomarkers will soon be discovered that, like *KDM6A* mutations, can predict tumour responsiveness to a particular therapy. This research avenue provides cause for optimism that improved outcomes for people with pancreatic cancer will be the norm — and not the exception — in the near future. ■

Fieke Froeling *and* **David Tuveson** *are at Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA.*
e-mails: froelin@cshl.edu; dtuveson@cshl.edu

1. Siegel, R. L., Miller, K. D. & Jemal, A. *CA Cancer J. Clinicians* **67**, 7–30 (2017).
2. Bailey, P. *et al. Nature* **531**, 47–52 (2016).
3. Collisson, E. A. *et al. Nature Med.* **17**, 500–503 (2011).
4. Moffitt, R. A. *et al. Nature Genet.* **47**, 1168–1178 (2015).
5. Andricovich, J. *et al. Cancer Cell* **33**, 512–526 (2018).
6. Shi, J. & Vakoc, C. R. *Mol. Cell* **54**, 728–736 (2014).
7. McDonald, O. G. *et al. Nature Genet.* **49**, 367–376 (2017).
8. Roe, J.-S. *et al. Cell* **170**, 875–888 (2017).
9. The Cancer Genome Atlas Research Network. *Cancer Cell* **32**, 185–203 (2017).
10. Waddell, N. *et al. Nature* **518**, 495–501 (2015).
11. Le, D. T. *et al. N. Engl. J. Med.* **372**, 2509–2520 (2015).
12. Mazur, P. K. *et al. Nature Med.* **21**, 1163–1171 (2015).

This article was published online on 16 May 2018.

CONDENSED-MATTER PHYSICS

Plasmon propagation pushed to the limit

Excitations called plasmons have the potential to miniaturize photonic devices, but are often short-lived. Microscopy reveals that plasmons in the material graphene can overcome this limitation at low temperatures. SEE LETTER P.530

JUSTIN C. W. SONG

Light can be confined and steered at the nanoscale using collective oscillations of electrons known as plasmons. But just as death and taxes are the only certainties in life, energy loss is the only certainty in plasmonics. The tighter the confinement of light, the shorter the lifetime of the plasmons[1] — a trade-off that is a major hurdle in the practical use of these oscillations. On page 530, Ni *et al.*[2] use a technique called scanning near-field optical microscopy to study plasmons in a single layer of carbon atoms known as graphene,

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

at cryogenic temperatures (60 kelvin). The authors show that the plasmons can produce extremely compact light confinement while retaining long lifetimes. They use their results to determine the fundamental limits of plasmon propagation in graphene.

The propagation of light involves the oscillation of electric and magnetic fields. This oscillation defines the relationship between the frequency and wavelength of light, and underpins the diffraction limit — the fact that, in free space, light spreads out if it passes through a region narrower than its wavelength. When light interacts with plasmons, its speed can be substantially reduced, which allows it to be confined to distances much smaller than its free-space wavelength. As a result, plasmons have become a versatile tool for controlling the behaviour of light at the nanoscale[3]. However, the same light–plasmon interaction that can confine light below its wavelength also enables energy to be lost through the scattering of electrons.

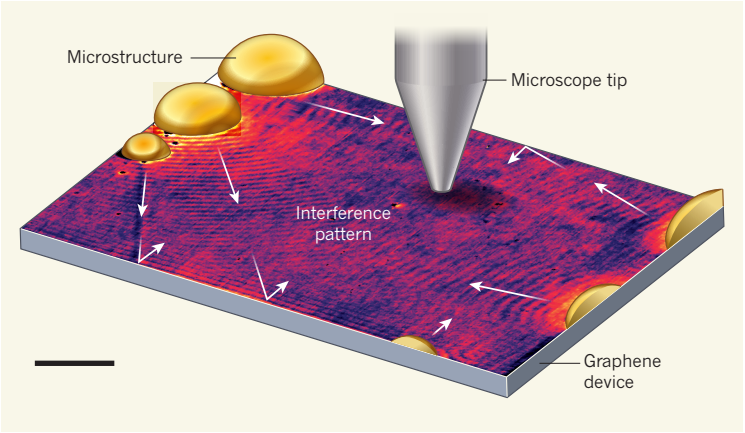


Figure 1 | Low-temperature plasmons investigated in graphene. Ni *et al.*[2] used an instrument known as a scanning near-field optical microscope to study exotic excitations called plasmons. They used the narrow metallic tip of the microscope to launch plasmons in a device containing the material graphene, at cryogenic temperatures (60 kelvin). The authors then scanned the tip across the device to image the interference pattern produced by the plasmons as they were reflected (white arrows) from the edges of the device and from microstructures present on the device's surface. The interference pattern consisted of bright and dark bands that were found throughout the device, demonstrating that the plasmons could travel several micrometres before their energy was lost. Such long-lived plasmons could have many applications. Scale bar, 1 μm. (Adapted from Fig. 1c of ref. 2.)

Noble metals such as silver and gold are conventionally used in plasmonics, but suffer from high losses. In the past few years, 2D materials have become promising alternatives[4]. In the case of graphene, plasmons can compress light to distances as small as one-three-hundredth of the light's free-space wavelength[5]. Furthermore, the electron density of graphene can be readily controlled, which provides direct electrical means of tuning the properties of the plasmons. But although sustained efforts to improve the quality of graphene have yielded steady advances[6], plasmon loss remains substantial.

In a bid to push the limits of plasmon propagation, Ni and colleagues launched and imaged plasmons in a device containing high-quality graphene at cryogenic temperatures. The use of these temperatures minimized losses caused by temperature-sensitive processes, such as the scattering of electrons from mechanical vibrations called phonons. The authors customized an instrument known as a scanning near-field optical microscope so that it could operate at cryogenic temperatures. Although these instruments are routinely used to study plasmons at room temperature, operating them at lower temperatures has been difficult.

Ni *et al.* used the narrow metallic tip of the microscope to launch plasmons in the graphene device. They then scanned the tip across the device to image the interference pattern produced by the plasmons as they reflected from the edges of the device and from microstructures present on the device's surface (Fig. 1). This technique is particularly useful because it launches plasmons in the interior of the device, which limits losses caused by interaction with the device's edges. Such losses can be large in other approaches[7].

The fruits of Ni and colleagues' labour are pronounced plasmon interference fringes (bright and dark bands) that are found throughout the device and that extend several micrometres from any boundaries. These fringes make the entire device 'light up' with a characteristic washboard-like pattern. The plasmons simultaneously have relatively long lifetimes (reaching 1.6 picoseconds; 1 ps is 10^{-12} seconds) and confine light to distances smaller than one-sixtieth of the free-space wavelength. Their quality factor, a measure of energy retention, is 130, which is a record for plasmons that enable such compact light confinement. The performance of the plasmons therefore bucks the trade-off between tight confinement and high loss. It is possible thanks to the extremely high quality of the authors' graphene device, which contains highly mobile electrons that can travel several micrometres without scattering.

Remarkably, using a combination of detailed modelling and systematically collected temperature-dependent data, Ni and colleagues determined that the main cause of energy loss at low temperatures was not electron scattering in the graphene. Instead, plasmon loss arose mostly from insulating material that surrounded the graphene. The quality of the plasmons could therefore be improved by reducing these extrinsic losses, for example by altering this insulating material. The authors also suggest that the intrinsic (fundamental) limits of plasmon propagation at cryogenic temperatures have not yet been reached. They calculate that it might be possible to achieve quality factors more than seven times higher than the one reported in the current paper.

Nevertheless, the exceptional quality of Ni and colleagues' graphene plasmons sets a new standard for nanophotonic platforms. Tightly confined light in such plasmons can now be thought of as being highly stable, with the ability to be directed and steered across distances of several micrometres. The possibilities for the future are vast and range from the fundamental (such as probing the topological[8] and geometrical structure[9] of plasmons) to the applied (including nanoscale plasmon lasers[10], sensitive light detectors, sub-wavelength routing of light, and nanoscale optical interconnects[3]). The authors' high-quality graphene plasmons, combined with recently developed techniques to substantially reduce the overall size of plasmons[11], make a compelling case for graphene-based nanophotonics.

Perhaps most exciting, however, is the prospect of using scanning near-field optical microscopy at cryogenic temperatures to probe excitations other than plasmons. Phases of matter such as superconductors, ferromagnets and antiferromagnets possess excitations that could be accessed using this technique[12]. In the past few years, a wide range of these phases has been discovered on 2D materials, on which the surfaces are fully exposed and are therefore easily accessible. Such phases manifest only at low temperatures, making cryogenic operation the key to launching the excitations and studying their intricate dynamics. ■

Justin C. W. Song *is in the Division of Physics and Applied Physics, Nanyang Technological University, Singapore 637371, Singapore, and the Institute of High Performance Computing, A*STAR, Singapore.*
e-mail: justinsong@ntu.edu.sg

1. Khurgin, J. B. *Nature Nanotech.* **10**, 2–6 (2015).
2. Ni, G. X. *et al. Nature* **557**, 530–533 (2018).
3. Atwater, H. A. *Sci. Am.* **296**, 56–63 (2007).
4. Boltasseva, A. & Atwater, H. A. *Science* **331**, 290–291 (2011).
5. Lundeberg, M. B. *et al. Science* **357**, 187–191 (2017).
6. Woessner, A. *et al. Nature Mater.* **14**, 421–425 (2015).
7. Yan, H. *et al. Nature Photon.* **7**, 394–399 (2013).
8. Jin, D. *et al. Nature Commun.* **7**, 13486 (2016).
9. Shi, L.-K. & Song, J. C. W. *Phys. Rev. X* **8**, 021020 (2018).
10. Bergman, D. J. & Stockman, M. I. *Phys. Rev. Lett.* **90**, 027402 (2003).
11. Alcaraz Iranzo, D. *et al. Science* **360**, 291–295 (2018).
12. Basov, D. N., Fogler, M. M. & García de Abajo, F. J. *Science* **354**, aag1992 (2016).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0124-0

Mutant phenotypes for thousands of bacterial genes of unknown function

Morgan N. Price[1], Kelly M. Wetmore[1], R. Jordan Waters[2], Mark Callaghan[1], Jayashree Ray[1], Hualan Liu[1], Jennifer V. Kuehl[1], Ryan A. Melnyk[1], Jacob S. Lamson[1], Yumi Suh[1], Hans K. Carlson[1], Zuelma Esquivel[1], Harini Sadeeshkumar[1], Romy Chakraborty[3], Grant M. Zane[4], Benjamin E. Rubin[5], Judy D. Wall[4], Axel Visel[2,6], James Bristow[2], Matthew J. Blow[2]*, Adam P. Arkin[1,7]* & Adam M. Deutschbauer[1,8]*

One-third of all protein-coding genes from bacterial genomes cannot be annotated with a function. Here, to investigate the functions of these genes, we present genome-wide mutant fitness data from 32 diverse bacteria across dozens of growth conditions. We identified mutant phenotypes for 11,779 protein-coding genes that had not been annotated with a specific function. Many genes could be associated with a specific condition because the gene affected fitness only in that condition, or with another gene in the same bacterium because they had similar mutant phenotypes. Of the poorly annotated genes, 2,316 had associations that have high confidence because they are conserved in other bacteria. By combining these conserved associations with comparative genomics, we identified putative DNA repair proteins; in addition, we propose specific functions for poorly annotated enzymes and transporters and for uncharacterized protein families. Our study demonstrates the scalability of microbial genetics and its utility for improving gene annotations.

Thousands of bacterial genomes have been sequenced, revealing the predicted amino acid sequences of millions of proteins. Only a small proportion of these proteins have been studied experimentally and the functions of most proteins have been predicted from their similarity to experimentally characterized proteins. However, about one-third of bacterial proteins are not similar enough to any characterized protein to be annotated by this approach[1]. Furthermore, these predictions are often incorrect, as homologous proteins may have different substrate specificities[2]. This sequence-to-function gap represents a growing challenge for microbiology, because new bacterial genomes are being sequenced at an ever-increasing rate, while experimental protein characterization continues to be relatively slow[1].

One approach for investigating the function of an unknown protein is to assess the consequences of a loss-of-function mutation of the corresponding gene under multiple conditions[3–6]. The mutant phenotypes can be combined with comparative genomics to provide evidence-based annotations for a fraction of the proteins[3,4]. Transposon mutagenesis followed by sequencing (TnSeq) measures mutant phenotypes genome-wide from a single experiment in which tens of thousands of different mutants are grown together[7,8]. Coupling TnSeq with random DNA barcoding of each mutant (RB-TnSeq) makes it easier to measure phenotypes across many conditions[9]. Here, we use RB-TnSeq to address the sequence-to-function gap by systematically exploring the mutant phenotypes of thousands of genes from each of 32 bacteria under multiple experimental conditions (Fig. 1a).

Mutant fitness compendia for 32 bacteria

To perform a systematic assessment of mutant phenotypes across a diverse set of bacteria, we studied 32 genetically tractable bacteria representing six different bacterial divisions and 23 different genera. In addition to 30 aerobic heterotrophs, we also studied a strictly anaerobic sulfate-reducing bacterium (*Desulfovibrio vulgaris*) and a strictly photosynthetic cyanobacterium (*Synechococcus elongatus*) (Fig. 1b). We generated a randomly barcoded transposon mutant library in each of the 32 bacteria, ten of which have been previously described[9–12]. For each mutant population, we used TnSeq to generate genome-wide maps of transposon insertion locations. Genes that have very few or no transposon insertions are likely to be essential for viability, or nearly so, in the conditions that were used to select the mutants (Supplementary Note 1). We identified 289–614 genes per bacterium that are likely to encode essential proteins (Fig. 1b; Supplementary Table 1).

To identify conditions for mutant fitness profiling, we tested the growth of the 30 aerobic heterotrophic bacteria in a range of conditions, including the utilization of 94 different carbon sources and 45 different nitrogen sources, and their inhibition by 34–55 stress compounds including antibiotics and metals (Supplementary Tables 2–4). In the typical mutant fitness experiment, we grew a pool of mutants for 4–8 generations and used DNA barcode sequencing[13] to compare the abundance of the mutants before and after growth (Fig. 1a). We defined gene fitness to be the $\log_2$ change in abundance of mutants in that gene during the experiment (Fig. 1a). For example, a gene fitness value of −2 means that the strains with transposon insertions in that gene dropped to 25% of their initial relative abundance by the end of the experiment, whereas a fitness value of 0 means that their relative abundance was unchanged. Genes with insufficient coverage were excluded from our data set ('no data' in Fig. 1b) and only protein-coding genes were considered. Including all replicates, we conducted a total of 4,870 genome-wide fitness experiments that met our criteria for biological and internal consistency[9], representing 26–129 experimental conditions for each bacterium (Fig. 1b, Supplementary Table 5).

To illustrate the consistency of our data with known protein functions, we examined fitness data for the three most common classes of experiments: carbon utilization, nitrogen utilization, and stress. For the utilization of D-fructose or 4-hydroxybenzoate as the sole source of carbon in *Cupriavidus basilensis*, the fitness data identified expected proteins for the catabolism of each substrate (Fig. 1c). Similarly, the

[1]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [2]Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [3]Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [4]Department of Biochemistry, University of Missouri, Columbia, MO, USA. [5]Division of Biological Sciences, University of California, San Diego, CA, USA. [6]School of Natural Sciences, University of California, Merced, CA, USA. [7]Department of Bioengineering, University of California, Berkeley, CA, USA. [8]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA. *e-mail: mjblow@lbl.gov; aparkin@lbl.gov; amdeutschbauer@lbl.gov

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

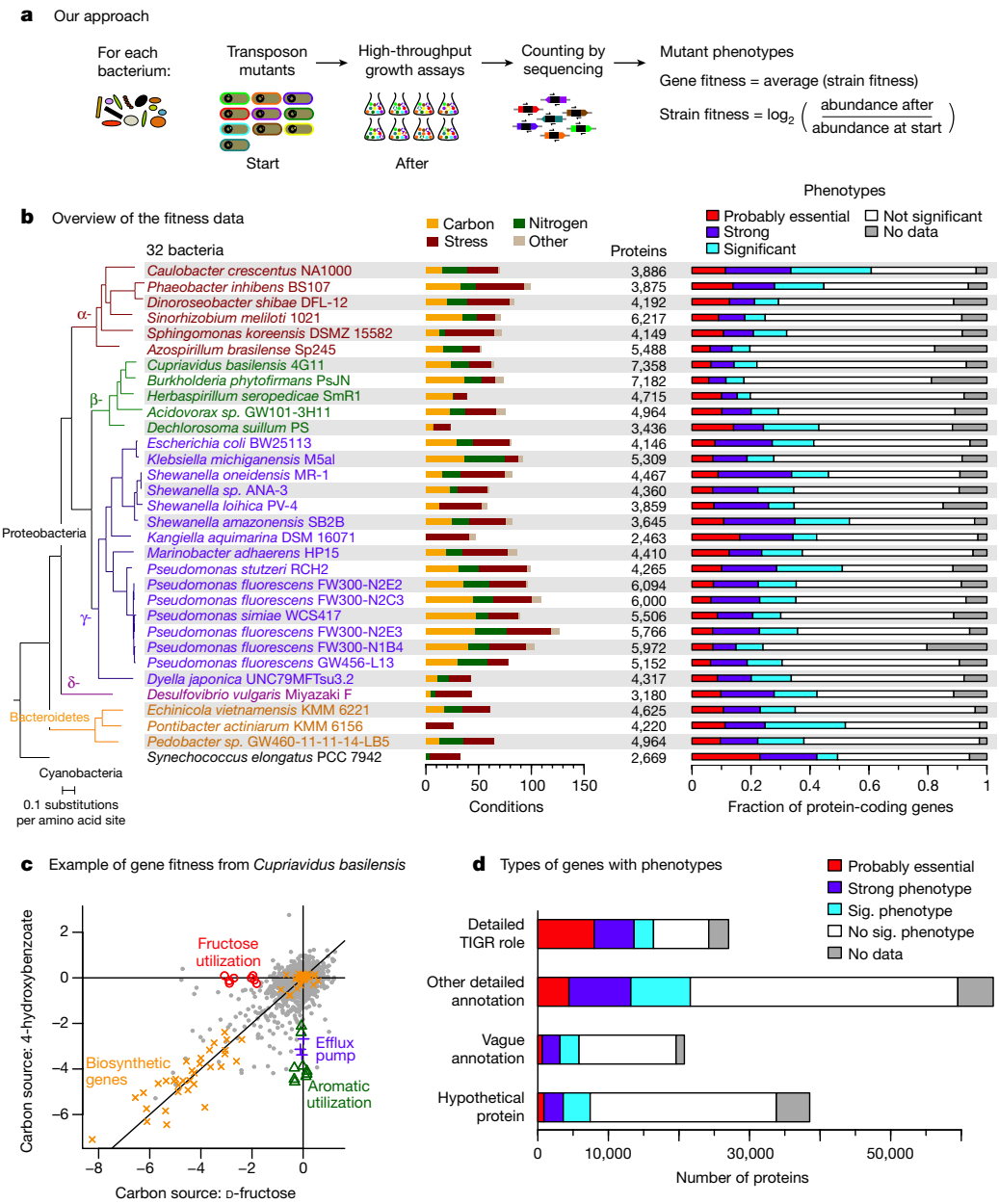


Fig. 1 | High-throughput genetics for 32 bacteria. a, Our approach for measuring gene fitness. **b**, For each bacterium, we show the types of conditions that we studied and how many genes had statistically significant mutant phenotypes (using a *t*-like test statistic[9] and FDR < 0.05). **c**, Gene fitness during the utilization of two carbon sources by *C. basilensis*. See Supplementary Table 6 for details on the highlighted genes. The 4-hydroxybenzoate data are the average from two biological replicates. **d**, We classified the genes from all 32 bacteria by how informative their annotations were, and for each class, we show how many genes have each type of phenotype.

fitness data identified the key enzymes and transporters required for the utilization of D-alanine or cytosine as the sole nitrogen source in *Azospirillum brasilense* (Extended Data Fig. 1a). Finally, in *Shewanella loihica*, orthologues of the CzcCBA heavy metal efflux pump[14] and the zinc-responsive regulator ZntR were important for fitness in the presence of an elevated concentration of zinc (Extended Data Fig. 1b). For each condition, we also identified proteins that were previously not known to be involved in the respective processes, including an efflux pump that is important for 4-hydroxybenzoate utilization by *C. basilensis* (Fig. 1c, Supplementary Table 6).

We computed a *t*-like test statistic for each gene fitness value[9] and identified a statistically significant mutant phenotype in at least one condition for 30% of the genes for which we collected fitness data (Fig. 1b). Eighteen per cent of all genes with fitness measurements were significantly detrimental to fitness (fitness > 0) in at least one condition (Extended Data Fig. 2), which is consistent with previous reports that many genes are detrimental in some conditions[3,15]. Genes annotated with a TIGRFAM[16] functional role (TIGR role) were particularly likely to have statistically significant phenotypes, with more than half of those with fitness data (52%) showing a significant phenotype (Fig. 1d). By contrast, genes with vague annotations that are not specific (such as 'transporter') or with functionally uninformative annotations (such as 'hypothetical protein' or 'membrane protein') were less likely to have phenotypes (28% or 20%, respectively). Nevertheless, our assays identified phenotypes for 11,779 genes that are not annotated with a detailed function (Fig. 1d), including 4,135 genes that encode proteins that do not belong to any characterized family in either Pfam or TIGRFAMs[16,17].

Conserved functional associations

To gain insight into the biological functions of individual proteins from the mutant fitness data, we used two strategies: (1) identification of

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

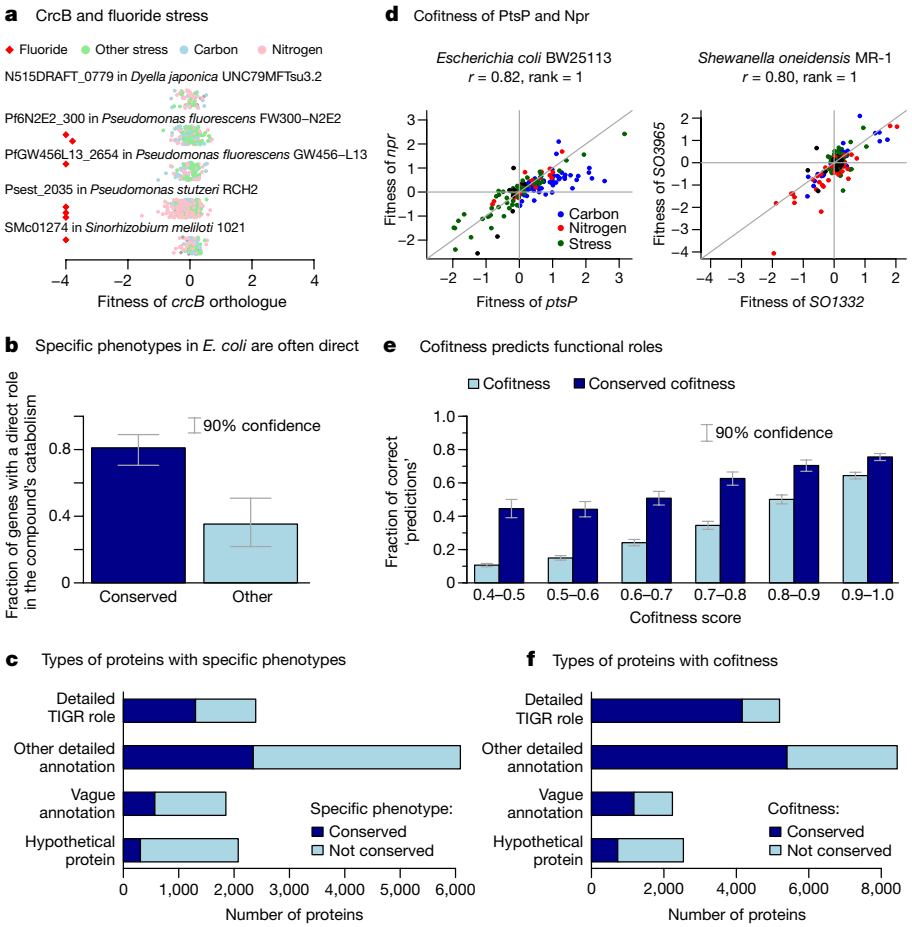


Fig. 2 | Identification of conserved phenotypes. a, An example of a conserved and specific phenotype. Each point shows the fitness of *crcB* in an experiment, with fluoride stress experiments highlighted in red. Values less than −4 are shown at −4. The *y*-axis is random. **b**, The fraction of *E. coli* genes with a specific phenotype in defined media whose encoded proteins are directly involved in the uptake or catabolism of the compound or regulation thereof. For this analysis, we examined all 61 *E. coli* genes with conserved specific-important phenotypes and a random sample of 40 of the other genes with specific-important phenotypes. The confidence interval is from the binomial test. **c**, The number of proteins of each type that had a conserved specific phenotype or a specific phenotype. **d**, Comparison of fitness values for *ptsP* and *npr* from *E. coli* ($n = 162$ independent experiments) and *S. oneidensis* ($n = 176$ independent experiments). The experiments are colour-coded by type. *r* is the linear correlation. **e**, Using TIGR subroles to test the accuracy of gene–gene associations. For each query gene with a TIGR subrole, we asked how often the most cofit gene with a TIGR subrole had the same subrole as the query gene, across varying levels of cofitness or conserved cofitness for that most cofit gene. The confidence interval is from the binomial test. **f**, The number of proteins of each type that had at least one association from conserved cofitness ($r > 0.6$ in both bacteria) or else from cofitness ($r > 0.8$).

'specific' phenotypes that are observed only under one or a small number of conditions in a given bacterium; and (2) 'cofitness' patterns, where multiple genes in a bacterium show similar fitness profiles across all conditions. Furthermore, we identified conserved specific phenotypes and conserved cofitness by comparing the data from 32 bacteria, and we tested the reliability of these conserved associations for understanding protein function.

To assign specific phenotypes, we identified genes that had $|\text{fitness}| > 1$ and $|t| > 5$ in an experiment but had little phenotype in most other experiments. For example, the fluoride efflux protein CrcB[18] is specifically important for fitness under elevated fluoride stress in eight bacteria, but not for fitness in any of the hundreds of other experimental conditions that we tested (Fig. 2a shows data from five bacteria). Among all genes with a significant phenotype under any condition, 33% had a specific phenotype. We considered a specific phenotype to be conserved if a similar protein in another bacterium (a putative orthologue) had a similar phenotype.

As catabolic processes in *Escherichia coli* are well understood, we compared the specific-important phenotypes (fitness < 0) on carbon and nitrogen sources to the EcoCyc database[19]. We found that many of these gene–condition associations resulted from the gene's direct involvement in the uptake or catabolism of the compound or in the regulation of those processes (Supplementary Table 7). If the specific phenotype was conserved, then the association was much more likely to be direct (Fig. 2b; $P < 10^{-4}$, Fisher's exact test).

We identified specific phenotypes and conserved specific phenotypes for genes of all annotation classes (Fig. 2c). In particular, specific phenotypes linked 3,927 genes with vague or hypothetical annotations to over 100 conditions, including 82 carbon sources, 43 nitrogen sources, and 54 stresses.

Our second strategy for gaining insight into a protein's function was based on the observation that genes with related functions often have similar fitness patterns across multiple conditions, which we term cofitness. Cofitness is the Pearson (linear) correlation of all of the fitness values for a pair of genes in the same bacterium, and cofitness across dozens of conditions has already been shown to be useful for understanding protein function[3–5,20]. For example, Npr and PtsP of the nitrogen phosphotransferase system (PTS) in *E. coli* were cofit (Fig. 2d), probably because PtsP phosphorylates and activates Npr[21]. We defined conserved cofitness as a pair of orthologous genes that had high cofitness in more than one bacterium, regardless of the conditions. For example, the orthologues of Npr and PtsP (SO1332 and SO3965) also have high cofitness in *Shewanella oneidensis*, but they have phenotypes in different conditions than in *E. coli* (Fig. 2d).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Phenotypic variability among orthologous genes has been reported before[22].

To test how accurately cofitness or conserved cofitness linked together functionally related genes, we determined for each query gene whether its functional role (TIGR subrole[16]) could be accurately predicted by the roles of its cofit genes. We found that high-scoring cofitness in a single bacterium led to predictions of TIGR subroles that were mostly correct, but the accuracy decayed rapidly as the cofitness score decreased (Fig. 2e). By contrast, for conserved cofitness, the decay was much slower (Fig. 2e). Furthermore, conserved cofitness was significantly more accurate for a given number of predictions: for example, the top 2,000 predictions from cofitness ($r > 0.86$ for gene pairs from one bacterium) had 63% agreement in TIGR subroles, while the top 2,000 predictions from conserved cofitness ($r > 0.66$ for gene pairs from both bacteria) had 74% agreement ($P < 10^{-12}$, Fisher's exact test). Conserved cofitness may be more predictive because it filters out cases of spurious cofitness between functionally unrelated genes in one bacterium. Using thresholds of $r > 0.8$ for cofitness or $r > 0.6$ for conserved cofitness, we identified at least one association from cofitness or conserved cofitness for 15% of the genes with fitness data and for 44% of genes with statistically significant phenotypes. We identified cofitness associations for all types of proteins (Fig. 2f), including for 4,773 vaguely annotated or hypothetical proteins.

Overall, we identified a functional association (either a specific phenotype or high cofitness) for 25,276 genes, of which 13,192 (52%) had conserved functional associations. Among the genes with

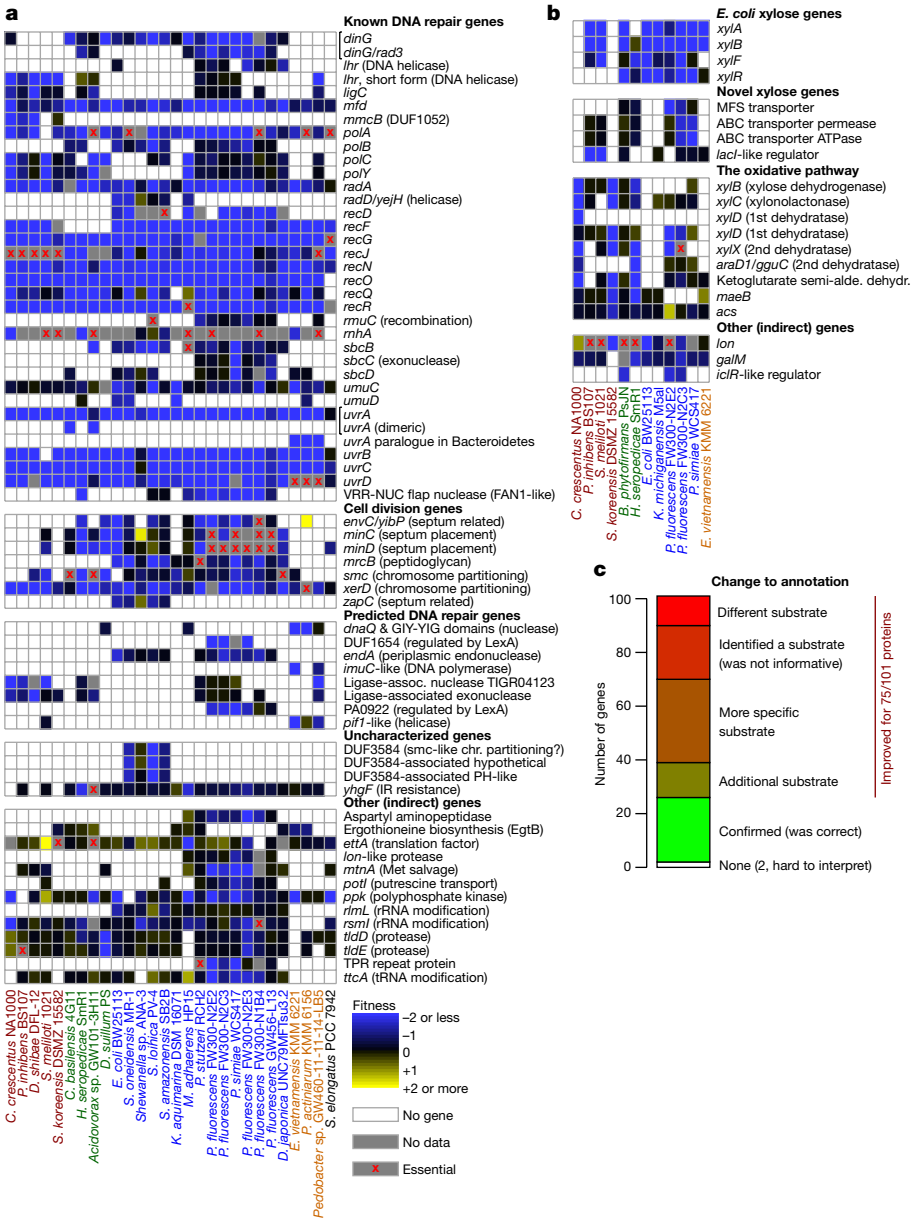


Fig. 3 | Genetic overviews for a condition or a class of proteins. **a**, Overview of conserved specific phenotypes in cisplatin stress across 28 bacteria. The data are the average for all successful cisplatin experiments for each bacterium, at up to five different concentrations. Each row shows an orthologue group formed by greedy clustering of orthologues (bidirectional best BLAST hits with 80% coverage). Some families are split into multiple orthologue groups and are marked with square brackets (*dinG*, *uvrA*). Some genes have more pleiotropic phenotypes in some (but not all) bacteria. **b**, Overview of specific phenotypes for the utilization of D-xylose as a carbon source in 12 bacteria. Putative orthologues are included in the heatmap even if they are not important for D-xylose utilization. In the oxidative pathway, gene names for *xylBCDX* are from Stephens et al.[33] and should not be confused with *E. coli xylB*, which is not related. The data are the average of 1–3 replicate experiments for each bacterium. The colour scale is the same as in **a**. **c**, Summary of annotation improvements for ABC transporter proteins based on an analysis of specific phenotypes.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

conserved associations, 10,699 (81%) had conserved associations across genera and 7,811 (59%) had conservation across divisions. For 2,316 genes with hypothetical or otherwise vague annotations, we identified conserved and hence high-confidence associations (Supplementary Table 8).

Genetic overviews of cellular processes

Genome-wide mutant fitness profiling of diverse bacteria provides a broad genetic overview of each biological condition studied. For example, cisplatin reacts with DNA to form crosslinks that block DNA replication, so we expected that genes encoding DNA repair proteins would be specifically important for growth during cisplatin stress[3]. Indeed, of the 67 protein families that had specific, important, and conserved phenotypes on cisplatin, 33 are known to be involved in DNA repair, including the UvrABC nucleotide excision repair complex (Fig. 3a, Supplementary Table 9). Three of these proteins have recently been shown to be involved in DNA repair: RadD (YejH)[23], MmcB (DUF1052)[24], and FAN1-like VRR-NUC domain protein[25], and individual mutants in these genes were sensitive to cisplatin (Extended Data Figs. 3, 4). Seven of the other characterized families that had conserved sensitivity to cisplatin are involved in cell division or chromosome segregation (Fig. 3a), probably because DNA damage can inhibit cell division and lead to filamentous cells[26]. Among the remaining 27 families, we predicted that eight poorly understood families were novel DNA repair families because of their domain content or because of regulation by the DNA damage response regulator LexA[27–29], including the nuclease EndA (Supplementary Note 2). An *endA* deletion strain is sensitive to cisplatin (Extended Data Fig. 4) and the catalytic residue of the nuclease domain of EndA is important for cisplatin resistance (Extended Data Fig. 5).

We obtained similar genetic overviews for many metabolic processes. For example, we examined D-xylose catabolism, which we assayed as the sole carbon source in 12 bacteria. We found that XylAB was important for xylose utilization in *E. coli* and in eight other bacteria, confirming its central and conserved role (Fig. 3b). By contrast, the well-characterized *E. coli* XylR regulator and XylF transporter are not required in each of the other eight bacteria: two Pseudomonads use alternative transport proteins for D-xylose, whereas *Phaeobacter inhibens* and *Sinorhizobium meliloti* require a LacI-like regulator for D-xylose utilization, as previously predicted for *P. inhibens* (PGA1_c13990)[9,30]. Three bacteria use an oxidative pathway for D-xylose utilization[31–33] instead of the XylAB pathway (Supplementary Table 10).

More accurate gene annotations

Large-scale mutant fitness data can also be used to improve our understanding of proteins that have been annotated with a general biochemical function but lack substrate specificity. To illustrate this, we used the mutant fitness data to systematically re-annotate the substrate specificities of 101 permease subunits of ABC transporters that had strong and specific-important phenotypes (fitness < -2) during the utilization of diverse carbon or nitrogen sources (Fig. 3c, Supplementary Table 11). Using the fitness data, we predicted substrates for all of these proteins (Supplementary Note 3). For 24 of the 50 ABC transport proteins that had already been annotated with a substrate (48%), the annotation was incorrect or did not include all of the substrates. Our data also provided more specific annotations: for example, Ac3H11_2942 and Ac3H11_2943 from *Acidovorax* sp. GW101-3H11 were annotated as transporting 'various polyols', whereas our data show that they are important for using the polyol D-sorbitol but not the polyol D-mannitol. Overall, we improved the annotations for 75 of 101 transport proteins (Fig. 3c).

Next we examined transport proteins with specific-important phenotypes in carbon or nitrogen source experiments and catabolic enzymes with specific-important phenotypes in carbon experiments, and identified instances in which the mutant fitness data led to a new annotation. In total, we re-annotated 456 proteins that were annotated vaguely or incorrectly in either KEGG[34] or SEED[35]: 238 transport proteins (this includes 68 of the ABC transport proteins described above) and 218 catabolic proteins (Supplementary Table 12). Of these 456 proteins, 287 (63%) were not annotated correctly by either KEGG or SEED. Most of the re-annotated proteins are homologous to characterized enzymes but were too distant for the correct substrate to be identified computationally. We also identified a number of proteins that we could link to novel enzymatic reactions (Supplementary Note 4, Supplementary Figs. 1–3). For example, we identified the putative *Pseudomonas* gene encoding glucosaminate ammonia-lyase, a known biochemical activity with no known gene[36].

Insights for uncharacterized families

We identified a conserved functional association for 335 genes that encode representatives of 87 different domains of unknown function (DUFs)[17] (Supplementary Table 13). We examined the phenotypes of

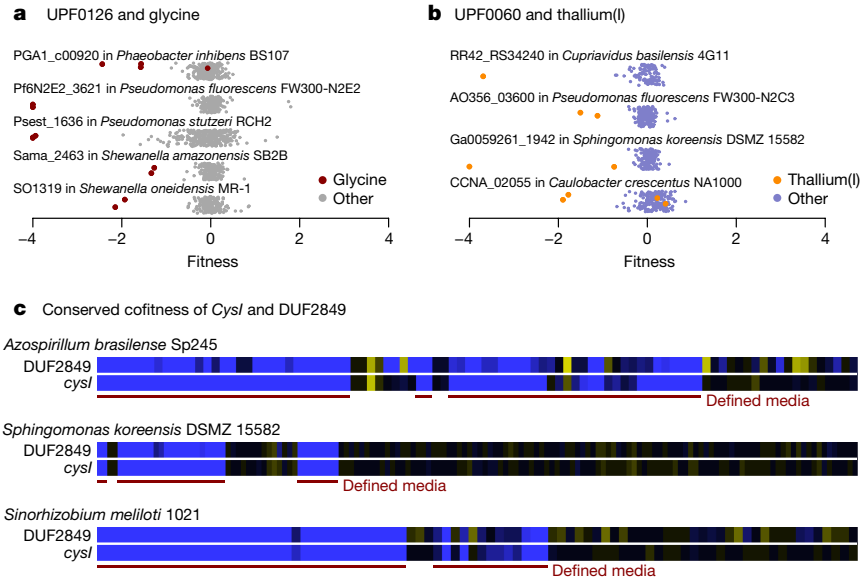


Fig. 4 | Conserved functional associations for genes encoding uncharacterized protein families. a, **b**, Conserved specific phenotypes for members of uncharacterized protein families UPF0126 and UPF0060. Each point represents the fitness of the protein in an individual experiment. Values under −4 are shown at −4. The *y*-axis is random. **c**, Heatmap of fitness data for *cysI* (sulfite reductase) and genes containing the domain of unknown function DUF2849 in three bacteria.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

77 of these DUFs, and we propose broad functional annotations for 13 of them and specific molecular functions for an additional eight (Supplementary Note 5). For example, proteins containing UPF0126 are specifically important for glycine utilization in 11 bacteria (Fig. 4a shows the data from five bacteria). As UPF0126 is predicted to be a membrane protein, we propose that it is a glycine transporter. Individual mutants of three members of this family had reduced growth on glycine (Extended Data Fig. 6) and PGA1_c00920 partially rescues the glycine growth defect of an *E. coli* strain that lacks the glycine transporter CycA[37] (Extended Data Fig. 7). Second, we found that genes encoding members of the UPF0060 family of predicted transmembrane proteins have a conserved specific phenotype under thallium stress in four bacteria (Fig. 4b). Consequently, we propose that UPF0060-containing proteins may function as a thallium-specific efflux pump. In support of this hypothesis, the expression of UPF0060 proteins confers thallium resistance on *E. coli* (Extended Data Fig. 8). Third, we found that in three bacteria, genes encoding DUF2849-containing proteins were cofit with an adjacently encoded sulfite reductase, CysI (Fig. 4c). Sulfite reductase is important for sulfate assimilation, which is the only source of sulfur in our defined medium. Thus, DUF2849 is also presumably involved in the same process. The three bacteria containing DUF2849 lack CysJ, which typically provides the electron source for CysI. As other bacterial genomes that contain DUF2849 also contain *cysI* but not *cysJ*, we propose that DUF2849 is an alternate electron source for sulfite reductase. Finally, we found that the uncharacterized proteins YeaH and YcgB and the poorly characterized protein kinase YeaG[38,39] had high cofitness across seven different bacteria, but with varying phenotypes across bacteria (Supplementary Figs. 4, 5). Given the protein kinase activity of YeaG, we propose that these three proteins act together in a conserved signalling pathway that is required for distinct cellular functions in different bacteria.

Discussion

We identified mutant phenotypes for more than ten thousand poorly annotated genes from 32 bacteria. By manually combining these functional associations with comparative sequence analysis, we proposed specific functions for transporter proteins, catabolic enzymes, and DUFs, and we identified putative novel DNA repair families. Most of these predictions require additional experimental validation. To facilitate further analyses of mutant phenotypes and protein sequences, we developed the Fitness Browser (http://fit.genomics.lbl.gov).

A major challenge in extending our results to all bacterial proteins is their incredible diversity. We identified functional associations for potential orthologues of just 12% of all bacterial proteins that lack detailed annotations (Supplementary Note 6, Extended Data Fig. 9). Improving this coverage will require a larger effort to generate mutants in more diverse bacteria: our study included representatives of only six of the approximately forty divisions of bacteria that have been cultivated so far. In summary, our study demonstrates the scale with which bacterial fitness data can be collected and the utility of these data to provide insights into the functions of many proteins.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0124-0.

Received: 5 October 2016; Accepted: 9 April 2018;
Published online: 16 May 2018

1. Chang, Y.-C. et al. COMBREX-DB: an experiment centered database of protein function: knowledge, predictions and knowledge gaps. *Nucleic Acids Res.* **44**, D330–D335 (2016).
2. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLOS Comput. Biol.* **5**, e1000605 (2009).
3. Deutschbauer, A. et al. Towards an informative mutant phenotype for every bacterial gene. *J. Bacteriol.* **196**, 3643–3655 (2014).
4. Deutschbauer, A. et al. Evidence-based annotation of gene function in *Shewanella oneidensis* MR-1 using genome-wide fitness profiling across 121 conditions. *PLoS Genet.* **7**, e1002385 (2011).
5. Nichols, R. J. et al. Phenotypic landscape of a bacterial cell. *Cell* **144**, 143–156 (2011).
6. Price, M. N. et al. The genetic basis of energy conservation in the sulfate-reducing bacterium *Desulfovibrio alaskensis* G20. *Front. Microbiol.* **5**, 577 (2014).
7. Langridge, G. C. et al. Simultaneous assay of every *Salmonella typhi* gene using one million transposon mutants. *Genome Res.* **19**, 2308–2316 (2009).
8. van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods* **6**, 767–772 (2009).
9. Wetmore, K. M. et al. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio* **6**, e00306–e00315 (2015).
10. Liu, H. et al. Magic pools: parallel assessment of transposon delivery vectors in bacteria. *mSystems* **3**, e00143–17 (2018).
11. Rubin, B. E. et al. The essential gene set of a photosynthetic organism. *Proc. Natl Acad. Sci. USA* **112**, E6634–E6643 (2015).
12. Melnyk, R. A. et al. Novel mechanism for scavenging of hypochlorite involving a periplasmic methionine-rich peptide and methionine sulfoxide reductase. *MBio* **6**, e00233–15 (2015).
13. Smith, A. M. et al. Quantitative phenotyping via deep barcode sequencing. *Genome Res.* **19**, 1836–1842 (2009).
14. Rensing, C., Pribyl, T. & Nies, D. H. New functions for the three subunits of the CzcCBA cation-proton antiporter. *J. Bacteriol.* **179**, 6871–6879 (1997).
15. Hottes, A. K. et al. Bacterial adaptation through loss of function. *PLoS Genet.* **9**, e1003617 (2013).
16. Haft, D. H. et al. TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.* **41**, D387–D395 (2013).
17. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
18. Baker, J. L. et al. Widespread genetic switches and toxicity resistance proteins for fluoride. *Science* **335**, 233–235 (2012).
19. Keseler, I. M. et al. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* **41**, D605–D612 (2013).
20. Hillenmeyer, M. E. et al. Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. *Genome Biol.* **11**, R30 (2010).
21. Rabus, R., Reizer, J., Paulsen, I. & Saier, M. H. Jr Enzyme I^{Ntr} from *Escherichia coli*. A novel enzyme of the phosphoenolpyruvate-dependent phosphotransferase system exhibiting strict specificity for its phosphoryl acceptor, NPr. *J. Biol. Chem.* **274**, 26185–26191 (1999).
22. van Opijnen, T., Dedrick, S. & Bento, J. Strain dependent genetic networks for antibiotic-sensitivity in a bacterial pathogen with a large pan-genome. *PLoS Pathog.* **12**, e1005869 (2016).
23. Chen, S. H., Byrne, R. T., Wood, E. A. & Cox, M. M. *Escherichia coli* radD (yejH) gene: a novel function involved in radiation resistance and double-strand break repair. *Mol. Microbiol.* **95**, 754–768 (2015).
24. Lopes-Kulishev, C. O. et al. Functional characterization of two SOS-regulated genes involved in mitomycin C resistance in *Caulobacter crescentus*. *DNA Repair (Amst.)* **33**, 78–89 (2015).
25. Gwon, G. H. et al. Crystal structure of a Fanconi anemia-associated nuclease homolog bound to 5′ flap DNA: basis of interstrand cross-link repair by FAN1. *Genes Dev.* **28**, 2276–2290 (2014).
26. Justice, S. S., Hunstad, D. A., Cegelski, L. & Hultgren, S. J. Morphological plasticity as a bacterial survival strategy. *Nat. Rev. Microbiol.* **6**, 162–168 (2008).
27. da Rocha, R. P., Paquola, A. C. de M., Marques Mdo, V., Menck, C. F. M. & Galhardo, R. S. Characterization of the SOS regulon of *Caulobacter crescentus*. *J. Bacteriol.* **190**, 1209–1218 (2008).
28. Abella, M., Campoy, S., Erill, I., Rojo, F. & Barbé, J. Cohabitation of two different lexA regulons in *Pseudomonas putida*. *J. Bacteriol.* **189**, 8855–8862 (2007).
29. Cirz, R. T., O'Neill, B. M., Hammond, J. A., Head, S. R. & Romesberg, F. E. Defining the *Pseudomonas aeruginosa* SOS response and its role in the global response to the antibiotic ciprofloxacin. *J. Bacteriol.* **188**, 7101–7110 (2006).
30. Wiegmann, K. et al. Carbohydrate catabolism in *Phaeobacter inhibens* DSM 17395, a member of the marine roseobacter clade. *Appl. Environ. Microbiol.* **80**, 4725–4737 (2014).
31. Brouns, S. J. J. et al. Identification of the missing links in prokaryotic pentose oxidation pathways: evidence for enzyme recruitment. *J. Biol. Chem.* **281**, 27378–27388 (2006).
32. Johnsen, U. et al. D-xylose degradation pathway in the halophilic archaeon *Haloferax volcanii*. *J. Biol. Chem.* **284**, 27290–27303 (2009).
33. Stephens, C. et al. Genetic analysis of a novel pathway for D-xylose metabolism in *Caulobacter crescentus*. *J. Bacteriol.* **189**, 2181–2185 (2007).
34. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
35. Overbeek, R. et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206–D214 (2014).
36. Iwamoto, R. & Imanaga, Y. Direct evidence of the Entner–Doudoroff pathway operating in the metabolism of d-glucosamine in bacteria. *J. Biochem.* **109**, 66–69 (1991).
37. Ghrist, A. C. & Stauffer, G. V. The *Escherichia coli* glycine transport system and its role in the regulation of the glycine cleavage enzyme system. *Microbiology* **141**, 133–140 (1995).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

38. Figueira, R. et al. Adaptation to sustained nitrogen starvation by *Escherichia coli* requires the eukaryote-like serine/threonine kinase YeaG. *Sci. Rep.* **5**, 17524 (2015).
39. Tagourti, J., Landoulsi, A. & Richarme, G. Cloning, expression, purification and characterization of the stress kinase YeaG from *Escherichia coli*. *Protein Expr. Purif.* **59**, 79–85 (2008).

Acknowledgements We thank V. Lo, W. Shao, and K. Keller for technical assistance with the Fitness Browser website. Sequencing was performed at: the Vincent J. Coates Genomics Sequencing Laboratory (University of California at Berkeley), supported by NIH S10 Instrumentation Grants S10RR029668, S10RR027303, and OD018174; the DOE Joint Genome Institute; the College of Biological Sciences UCDNA Sequencing Facility (UC Davis); and the Institute for Genomics Sciences (University of Maryland). Studies of novel isolates were conducted by ENIGMA and were supported by the Office of Science, Office of Biological and Environmental Research of the US Department of Energy, under contract DE-AC02-05CH11231. The other data collection was supported by Laboratory Directed Research and Development (LDRD) funding from Berkeley Laboratory, provided by the Director, Office of Science, of the US Department of Energy under contract DE-AC02-05CH11231 and a Community Science Project from the Joint Genome Institute to M.J.B., J.B., A.P.A., and A.M.D. The work conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231.

Author contributions A.M.D., A.P.A., M.N.P., M.J.B., and J.B. conceived the project. A.M.D., A.P.A., M.J.B., and J.B. supervised the project. A.M.D. led the experimental work. A.M.D., K.M.W., R.J.W., R.A.M., M.C., J.R., J.V.K., H.L., H.K.C., J.S.L., Y.S., Z.E., and H.S. collected data. R.C. isolated bacteria. M.N.P. and A.M.D. analysed the fitness data. R.A.M., R.J.W., and M.N.P. assembled genomes. B.E.R. provided resources and advice on *S. elongatus* experiments. G.M.Z. and J.D.W. generated gene deletion mutants in *Pseudomonas stutzeri* RCH2. A.V. edited the manuscript and provided advice. M.N.P., M.J.B., and A.M.D. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0124-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0124-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.J.B., A.P.A., & A.M.D.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Bacteria for high-throughput genetics. We attempted to make transposon mutant libraries in more than 100 bacteria as part of this study, including representatives of Proteobacteria, Bacteroidetes, Firmicutes, Actinobacteria, and Planctomycetes; we present data from the 32 bacteria to which we successfully applied RB-TnSeq (Supplementary Table 14). Eight bacteria were isolated from groundwater collected from different monitoring wells at the Oak Ridge National Laboratory Field Research Center (FRC; https://public.ornl.gov/orifc/orfrc1_fieldchallenge.cfm), and five have not been described previously: *Acidovorax* sp. GW101-3H11, *Pseudomonas fluorescens* FW300-N1B4, *P. fluorescens* FW300-N2E3, *P. fluorescens* FW300-N2C3, and *P. fluorescens* GW456-L13. *Acidovorax* sp. GW101-3H11 was isolated as a single colony on a Luria-Bertani (LB) agar plate grown at 30 °C using an inoculum from FRC well GW101. *P. fluorescens* FW300-N1B4, *P. fluorescens* FW300-N2E3, and *P. fluorescens* FW300-N2C3 were all isolated at 30 °C under anaerobic denitrifying conditions with acetate, propionate, and butyrate as the carbon source, respectively, using an inoculum from FRC well FW300. *P. fluorescens* GW456-L13 was isolated from FRC well FW456 under anaerobic incubations on an LB agar plate. We previously described the isolation of *P. fluorescens* FW300-N2E2[40], *C. basilensis* 4G11[41], and *Pedobacter* sp. GW460-11-11-14-LB5[10]. The eight FRC isolates have been submitted to the Leibniz Institute DSMZ (German Collection of Microorganisms and Cell Cultures GmbH).

Our study included seven strains of *Pseudomonas* and four strains of *Shewanella* because of their genetic tractability. The different strains in these genera have large differences in gene content. For example, the two most closely related strains we studied were *P. fluorescens* FW300-N2E2 and FW-N2C3, and the typical pair of orthologous proteins from these strains has 97% amino acid identity. Nevertheless, each genome contains over 1,000 predicted protein-coding genes that do not have orthologues in the other strain.

Strain construction for single gene studies. The oligonucleotides and gBlocks used in this study are listed in Supplementary Table 15, the plasmids and details on their construction in Supplementary Table 16, and the single strains for follow-up studies in Supplementary Table 17. We constructed kanamycin-marked gene deletions in *Pseudomonas stutzeri* RCH2 using a previously described double homologous recombination strategy[42]. We used a similar strategy to construct kanamycin-marked deletions in *Dinoroseobacter shibae* and *P. inhibens*, except that we delivered the deletion constructs by conjugation from *E. coli*. We constructed markerless gene deletions of SO4008 and SO1319 from *S. oneidensis* MR-1 and Pf6N2E2_4800 and Pf6N2E2_4801 from *P. fluorescens* FW300-N2E2 using *sacB* counter-selection. For genetic complementation experiments, we cloned the genes into the broad range vector pBBR1-MCS5[43]. For the thallium resistance experiments, we cloned members of UPF0060 into plasmid pFAB2286. For *E. coli* BW25113, we used single-gene deletions from the Keio collection[44]. For *S. oneidensis* MR-1, we also used transposon mutants that had been individually sequenced[4]. Unless indicated otherwise, we tested the growth phenotypes of these individual strains in 96-well microplates at 30 °C using the standard rich or minimal defined growth medium for each bacterium. These growth assays were performed in a Tecan microplate reader (either Sunrise or Infinite F200) with absorbance readings (OD_{600}) every 15 min.

Media and standard culturing conditions. A full list of the media used in this study and their components is given in Supplementary Table 18. We routinely cultured *Acidovorax* sp. GW101-3H11, *A. brasilense* Sp245, *Burkholderia phytofirmans* PsJN, *Dyella japonica* UNC79MFTsu3.2, *E. coli* BW25113, *Herbaspirillum seropedicae* SmR1, *Klebsiella michiganensis* M5a1, all Pseudomonads and Shewanellae, *S. meliloti* 1021, and *Sphingomonas koreensis* DSMZ 15582 in LB. *Caulobacter crescentus* NA1000 was typically cultured in PYE medium. *C. basilensis* 4G11 and *Pedobacter* sp. GW460-11-11-14-LB5 were typically cultured in R2A medium. *Dechlorosoma suillum* PS was cultured in ALP medium[12]. *D. vulgaris* Miyazaki F was grown anaerobically in lactate-sulfate (MOLS4) medium, as previously described[45,46]. We used marine broth (Difco 2216) for standard culturing of *D. shibae* DFL-12, *Echinicola vietnamensis*, *Kangiella aquimarina* SW-154T, *Marinobacter adhaerens* HP15, *P. inhibens* BS107, and *Pontibacter actiniarum*. *S. elongatus* PCC 7942 was normally cultured in BG-11 medium with either 7,000 or 9,250 lx. All bacteria were typically cultured at 30 °C except *E. coli* BW25113 and *Shewanella amazonensis* SB2B, which were cultured at 37 °C, and *P. inhibens* BS107, which was grown at 25 °C. The *E. coli* conjugation strain WM3064 was cultured in LB medium at 37 °C with diaminopimelic acid (DAP) added to a final concentration of 300 μM.

High-throughput growth assays of wild-type bacteria. To assess the phenotypic capabilities of 30 aerobic heterotrophic bacteria and to identify conditions suitable for mutant fitness profiling, we monitored the growth of the wild-type bacteria in a 96-well microplate assay. These prescreen growth assays were performed in a Tecan microplate reader (either Sunrise or Infinite F200) with absorbance readings (OD_{600}) every 15 min. All 96-well microplate growth assays contained 150 μl culture volume per well at a starting OD_{600} of 0.02. We used the grofit package in R[47] to analyse all growth curve data in this study. For carbon and nitrogen source utilization, we tested 94 and 45 possible substrates, respectively, in a defined medium (Supplementary Tables 2, 3). We classified a bacterium as positive for usage of a particular substrate if (1) the maximum OD_{600} on the substrate was at least 1.5 × greater than the average of the water controls and the integral under the curve (spline.integral) was 10% greater than the average of the water controls or (2) a successful genome-wide fitness assay was collected on the substrate, as described below. We included the second criterion because our automated scoring of the wild-type growth curves was conservative and did not include all conditions used for genome-wide fitness assays.

Additionally, for each of the 30 heterotrophic bacteria, we determined the inhibitory concentrations for 34–55 diverse stress compounds including antibiotics, biocides, metals, furans, aldehydes, and oxyanions. For each compound, we grew the wild-type bacterium across a 1,000-fold range of inhibitor concentrations in a rich medium. We used the spline.integral parameter of grofit to fit dose–response curves and calculate the half-maximum inhibitory concentration (IC_{50}) values for each compound (Supplementary Table 4). For *D. vulgaris* Miyazaki F and *S. elongatus* PCC 7942, we did not perform these growth prescreen assays; rather, we just performed the mutant fitness assays across a broad range of inhibitor concentrations.

Genome sequencing. We sequenced *Acidovorax sp*. GW101-3H11 and five Pseudomonads by using a combination of Illumina and Pacific Biosciences. For Illumina-first assembly, we used scythe (https://github.com/vsbuffalo/scythe) and sickle (https://github.com/najoshi/sickle) to trim and clean Illumina reads, we assembled with SPAdes 3.0[48], we performed hybrid Illumina/PacBio assembly on SMRTportal using AHA[49], we used BridgeMapper on SMRTportal to fix misassembled contigs, we mapped Illumina reads to the new assembly with bowtie 2[50], and we used pilon to correct local errors[51]. For *Acidovorax sp*. GW101-3H11, we instead used A5[52] to assemble the Illumina reads and we used AHA to join contigs together. For PacBio-first assembly, we used HGAP3 on SMRTportal, we used circlator to find additional joins[53], and we again used bowtie 2 and pilon to correct local errors. See Supplementary Table 19 for a summary of these genome assemblies and their accession numbers. In addition, *S. koreensis* DSMZ 15582 was sequenced for this project by the Joint Genome Institute, using Pacific Biosciences.

Constructing pools of randomly barcoded transposon mutants. The transposon mutant libraries for ten bacteria have been described previously[9–12]. The other 22 bacteria were mutagenized with randomly barcoded plasmids containing a *mariner* or *Tn*5 transposon, a *pir*-dependent conditional origin of replication, and a kanamycin resistance marker, using previously described vectors[9]. The plasmids were delivered by conjugation with *E. coli* WM3064, which is a diaminopimelate auxotroph and is pir^+. The conditions for mutagenizing each organism are described in Supplementary Table 20. Generally, we conjugated mid-log-phase grown WM3064 donor (either *mariner* donor plasmid library APA752 or *Tn*5 donor plasmid library APA766) and recipient cells on 0.45 μM nitrocellulose filters (Millipore) overlaid on rich medium agar plates supplemented with DAP. We used the rich medium preferred by the recipient (Supplementary Table 20). After conjugation, filters were resuspended in recipient rich medium and plated on recipient rich medium agar plates supplemented with kanamycin. After growth, we scraped together kanamycin-resistant colonies into recipient rich medium with kanamycin, diluted the culture back to a starting OD_{600} of 0.2 in 50–100 ml of recipient rich medium with kanamycin, and grew the mutant library to a final OD_{600} of between 1.0 and 2.0. We added glycerol to a final volume of 10%, made multiple 1- or 2-ml −80 °C freezer stocks, and collected cell pellets to extract genomic DNA for TnSeq. For *D. vulgaris* Miyazaki F, we selected for G418-resistant transposon mutants in liquid medium with no plating step (Supplementary Table 20).

Transposon insertion site sequencing (TnSeq). Given a pool of mutants, we performed TnSeq to amplify and sequence the transposon junction and to link the barcodes to a location in the genome[9]. In this study, we used the TnSeq data for two independent analyses. First, we used the mapped transposon insertions to identify genes that are likely to be essential for viability (or nearly so) in the conditions that we used to select the mutant library (see below). We performed this gene essentiality analysis independently of the DNA barcodes and we therefore considered transposon insertion locations that were not included in the mutant pool definition for BarSeq fitness assays. Second, we used the TnSeq data to define the mutant library for high-throughput mutant fitness assays. Here, we analysed both the transposon insertion data and its associated random DNA barcode, as the RB-TnSeq approach requires accurate association of the genomic insertion location of the transposon to a unique, random DNA barcode. We considered a barcode to be confidently mapped to a location if this mapping was supported by at least ten reads and the barcode mapped primarily to one location[9]. (For *Shewanella* sp. ANA-3, the threshold was eight reads.) The number of unique barcodes (strains) mapped in each mutant library is shown in Supplementary Table 20. Given this mapping, the abundance of the strains in each sample can be determined by a simpler and cheaper protocol: amplifying the barcodes with PCR followed by barcode sequencing[13].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Identifying essential or nearly essential genes. Genes that lack insertions or that have very low coverage in the start samples are likely to be essential or important for growth (nearly essential) in rich medium, as except for *S. elongatus*, pools of mutants were produced and recovered in medium that contained yeast extract. We used previously published heuristics[11] to distinguish likely-essential genes from genes that are too short or that are too repetitive to map insertions in. Briefly, for each protein-coding gene, we computed the total read density in TnSeq (reads/nucleotides across the entire gene) and the density of insertion sites within the central 10–90% of each gene (sites/nucleotides). We did not consider the DNA barcodes in this analysis of essential genes. Across the 32 bacteria, we mapped 5–66 insertions for the typical (median) protein-coding gene. We then excluded genes that might be difficult to map insertions within because they were very similar to other parts of the genome (BLAT score above 50) and also very short genes of less than 100 nucleotides. Given the median insertion density and the median length of the remaining genes, we investigated how short a gene could be and still be unlikely to have no insertions at all by chance ($P < 0.02$, Poisson distribution). Genes shorter than this threshold were excluded; the threshold varied from 100 nucleotides for *P. inhibens*, *C. crescentus*, and *S. elongatus* to 675 nucleotides for *S. loihica* PV-4. For the remaining genes, we normalized the read density by GC content by dividing by the running median of read density over a window of 201 genes (sorted by GC content). We normalized the insertion density so that the median gene's value was 1. Protein-coding genes were considered essential or important for growth (nearly essential) if we did not estimate fitness values for the gene and both the normalized insertion density and the normalized read density were under 0.2. A validation of this approach is described in Supplementary Note 1.

Mutant fitness assays. For each mutant library, we performed competitive mutant fitness assays under a large number of growth conditions that were chosen on the basis of the results of high-throughput growth assays of wild-type bacteria (see above). For some bacteria, we also assayed growth at varying pH or temperature, motility on agar plates, or survival. The conditions that we profiled varied across the bacteria owing to the differing growth capabilities of the bacteria as well as experimental limitations. For example, the heterotrophic bacteria we investigated showed a wide range in the number of compounds (either carbon or nitrogen) that they could utilize for growth in a defined medium (Supplementary Table 2, Supplementary Table 3). For example, only 14 of the 32 bacteria were capable of using D-xylose as the sole carbon source, and we successfully assayed mutant fitness in 12 of these strains. We did not perform carbon and nitrogen source experiments in *K. aquimarina* SW-154T and *P. actiniarum*, as we could not culture these bacteria in a defined growth medium. In addition, some stress experiments were not performed because of the native resistance of the bacteria to the compound (Supplementary Table 4).

In addition to biological reasons for the differences among the exact set of conditions we profiled, for some bacteria we did not attempt certain assays. We did not systematically test *D. vulgaris* Miyazaki F for carbon source utilization and we did not perform nitrogen source experiments in two heterotrophic bacteria that grow in defined medium: *D. suillum* PS and *S. loihica* PV-4. For 31 of the 32 bacteria, we attempted fitness assays for a core set of 32 stress compounds (see Supplementary Table 4), but for *D. suillum* PS, we attempted only 16 of them. We successfully studied motility in 12 of the 32 bacteria using a soft agar assay. For the other bacteria, we either did not attempt a motility assay or the cells were motile but the mutant fitness data did not pass our thresholds for a successful experiment. Similarly, there were many growth-based fitness experiments (carbon and nitrogen source, and stressors) that we performed but did not pass our metrics for a successful experiment[9]. Although a few dozen of these samples had low read depth, which might indicate a problem during PCR of the barcodes or sequencing, we believe that most of these experiments failed owing to biological factors that make them incompatible with a pooled fitness assay, such as intense positive selection or potentially stochastic exit from lag phase[9]. For example, across all 32 bacteria, there are 197 bacterium–stress combinations that lack fitness data because none of the experiments for that stress succeeded by our metrics. For 85 of these combinations, we attempted more than one experiment.

The full list of experiments performed for each mutant library along with detailed metadata are available in Supplementary Table 5, on Figshare (https://doi.org/10.6084/m9.figshare.5134837) and at http://genomics.lbl.gov/supplemental/bigfit/. Our analysis includes 385 successful experiments from Wetmore et al.[9] and 36 successful experiments from Melnyk et al.[12]. The other 4,449 successful fitness assays are described here for the first time. In general, all growth assays with carbon sources, nitrogen sources, and inhibitors were done as previously described[9]. In brief, an aliquot of the mutant library was thawed and inoculated into 25 ml rich medium with kanamycin or erythromycin and grown to mid-log phase in a flask. The only exception was *K. michiganensis*, which we recovered to stationary phase; experiments with the usual exponential phase start samples for this library had suspicious correlations of gene fitness with GC content and did not meet our quality metrics. Depending on the mutant library, this growth recovery took between 3 and 24 h. After recovery, we collected pellets for genomic DNA extraction and barcode sequencing (BarSeq) of the start sample. We used the remaining cells to set up multiple mutant fitness assays with diverse carbon and nitrogen sources in defined media and diverse inhibitors in rich media, all at a starting OD_{600} of 0.02. In addition, for most bacteria, we profiled growth of the mutant library at different pH values and temperatures. After the mutant library grew to saturation under the selective growth condition (typically 4–8 population doublings), we collected a cell pellet for genomic DNA extraction and BarSeq of the 'end' sample. As described below, we calculate gene fitness from the barcode counts of the end sample relative to the start sample.

We used a number of different growth formats and media formations for mutant fitness assays across the 32 bacteria. A full list of compound components for each growth medium are contained in Supplementary Table 18. Many fitness assays were done in 48-well microplates (Greiner) with 700 µl culture volume per well and grown in a Tecan Infinite F200 plate reader with OD_{600} measurements every 15 min. For these 48-well microplate assays, we combined the cultures from two replicate wells before genomic DNA extraction (total volume of experiment, 1.4 ml). For 24-well microplate experiments, we typically used deep-well plates with 1.5 ml (for inhibitors) or 2 ml (for carbon and nitrogen sources) total culture volume per well. For six bacteria (*C. crescentus*, *D. japonica* UNC79MFTsu3.2, *E. vietnamensis*, *K. michiganensis*, *Pedobacter* sp. GW460-11-11-14-LB5 and *P. actiniarum*), we performed all fitness assays in transparent 24-well microplates (Greiner) with 1.2 ml total volume per well. All 24-well microplate experiments were grown in a Multitron incubating shaker (Innova). For 24-well microplate experiments, we typically took the OD_{600} of each culture every 12–24 h in a Tecan plate reader (after transferring the cells to a Greiner 96-well microplate for the 24-well deep well plate experiments). Because we used different methods for measuring OD_{600} values (cuvettes with a spectrophotometer; 24, 48 and 96 well microplates in a plate reader), we used standard curves to interconvert microplate OD_{600} values to a common reference (cuvette with spectrophotometer). Over 1,000 experiments, primarily carbon source and temperature experiments, were done in glass test tubes with 5 ml culture volumes. For the test tube experiments, we monitored OD_{600} every 12–24 h with a standard spectrophotometer and cuvettes.

For stress experiments, we aimed to use an inhibitory but sub-lethal concentration of the compound, typically at a concentration that resulted in a 50% reduction in growth rate (IC_{50}). Both the growth and the inhibition are crucial for identifying mutants that have altered sensitivity to the compound. If the concentration of inhibitor is too high and there is no growth, then the abundance of the strains will not change and all of the fitness values will be near zero. If the concentration of inhibitor is too low and there is no growth inhibition, then the fitness pattern is likely to be as if the compound were not added at all. We identified these inhibitory concentrations using the growth curves with the wild-type bacteria described above. In addition to the calculated IC_{50} for a compound, we often used a few different concentrations above and below the IC_{50} to try to capture an inhibitory but sub-lethal concentration. We used multiple concentrations of a single compound because we found that the IC_{50} values determined with wild-type bacteria in 96-well microplates (see above) sometimes did not agree with the mutant fitness experiments, which were done with a complex transposon mutant library in 24 or 48-well microplates. For assays done in 48-well microplates and grown in a Tecan Infinite F200 plate reader, we could confirm that the culture was inhibited relative to a no-stress control. For stress assays in 24-well microplates and grown in the Multitron shaker, we took OD readings approximately every 12 h to estimate which cultures were inhibited. In practice, for a given mutant library, we often collected mutant fitness data with different concentrations of the same inhibitor. We also collected fitness data in plain rich medium without an added inhibitory compound.

For some carbon source experiments in *D. vulgaris* Miyazaki F, which is strictly anaerobic, we grew the mutant pool in 18 × 150-mm hungate tubes with a butyl rubber stopper and an aluminium crimp seal (Chemglass Life Sciences) with a culture volume of 10 ml and a headspace of about 15 ml. For the remainder of the *D. vulgaris* fitness experiments, we grew the mutant pool in 24-well microplates inside an anaerobic chamber. We used OD_{600} measurements to determine which cultures were inhibited by varying concentrations of stress compounds. Similarly, for six of the other heterotrophs, we measured gene fitness during anaerobic growth. All anaerobic media were prepared within a Coy anaerobic chamber with an atmosphere of about 2% H_2, 5% CO_2, and 93% N_2.

For *S. elongatus* PCC 7942, which is strictly photosynthetic, we recovered the library from the freezer in BG-11 medium at a light level of 7,000 lx and we conducted fitness assays at 9,250 lx. We used OD_{750} to measure the growth of *S. elongatus*. Most *S. elongatus* mutant fitness assays were done in the wells of a 12-well microplate (Falcon) with a 5 ml culture volume.

For motility assays, the mutant pool was inoculated into the centre of a 0.3% agar rich medium plate and 'outer' samples with motile cells were removed with a razor after 24–48 h. In many instances, we also removed an 'inner' sample of cells from near the point of inoculation. Not all bacteria we assayed were motile in this

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

soft agar assay and others were motile but did not give mutant fitness results that passed our quality metrics.

In four bacteria, we also assayed survival. In these assays, a mutant pool was subjected to a stressful condition (either extended stationary phase, starvation, or a low temperature of 4 °C) for a defined period; then, to determine which strains were still viable, they were recovered in rich medium for a few generations. After recovery in rich medium, the cells were harvested for genomic DNA extraction and BarSeq.

We performed replicate experiments (not necessary at the same concentration) for 25 of the 29 bacteria with carbon source data; the exceptions are *B. phytofirmans* PsJN, *P. fluorescens* FW300-N1B4, *P. fluorescens* FW300-N2E3, and *S. meliloti* 1021. Similarly, we performed replicate experiments (not necessarily at the same concentration) for 17 of the 28 bacteria with nitrogen source data. The bacteria with nitrogen source data but without biological replicates are: *Acidovorax* sp. GW101-3H11, *B. phytofirmans* PsJN, *C. basilensis* 4G11, *D. japonica* UNC79MFTsu3.2, *H. seropedicae* SmR1, *P. fluorescens* FW300-N1B4, *P. fluorescens* FW300-N2E3, *P. fluorescens* FW300-N2C3, *P. fluorescens* GW456-L13, *S. meliloti* 1021, and *P. simiae* WCS417. Lastly, we performed replicates (not necessarily at the same concentration) for the majority of stress experiments for 16 of the 32 bacteria: *C. crescentus* N1000, *E. vietnamensis*, *D. shibae* DFL-12, *K. aquimarina* SW-154T, *S. koreensis* DSMZ 15582, *K. michiganensis* M5a1, *M. adhaerens* HP15, *D. vulgaris* Miyazaki F, *S. oneidensis* MR-1, *P. inhibens* BS107, *P. actiniarum*, *D. suillum* PS, *P. stutzeri* RCH2, *Pedobacter* sp. GW460-11-11-14-LB5, *S. loihica* PV4, and *S. elongatus* PCC 7942.

Barcode sequencing (BarSeq). Genomic DNA extraction and barcode PCR were performed as described previously[9]. Most genomic DNA extractions were done in a 96-well format using a QIAcube HT liquid handling robot (QIAGEN). We used the 98 °C BarSeq PCR protocol[9], which is less sensitive to high GC content than the original 95 °C protocol. In general, we multiplexed 48 samples or 96 samples per lane of Illumina HiSeq. For *E. coli*, we sequenced 96 samples per lane. For sequencing runs on the Illumina HiSeq4000, we sequenced 96 samples per lane. For the HiSeq4000 runs, we used an equimolar mixture of four common P1 oligos for BarSeq, with variable lengths of random bases at the start of the sequencing reactions (2–5 nucleotides) (Supplementary Table 15). We did this to phase the amplicons for sequencing and to aid in cluster discrimination on the HiSeq4000.

Computation of fitness values. Fitness data was analysed as previously described[9]. In brief, the fitness value of each strain (an individual transposon mutant) is the normalized $\log_2$(strain barcode abundance at end of experiment/strain barcode abundance at start of experiment). The fitness value of each gene is the weighted average of the fitness of its strains. In this study, we restricted our analysis to the 123,255 different non-essential protein-coding genes for which we collected gene fitness data. Gene fitness values describe the relative abundance of the mutant strains in a condition, regardless of their fitness in rich medium. Only strains containing insertions within the central 10–90% of the gene and with sufficient abundance, on average, in the start samples were included in these calculations; we used 3–26 mutant strains for the typical protein-coding gene in each bacterium (Supplementary Table 20). Specifically, in the average start samples, we required that each gene have at least 30 reads, and each individual strain for that gene to have at least 3 reads. Because we include the same set of strains in the analysis of each experiment, there are no instances where a gene has a fitness value in one experiment but not in another. The gene fitness values were then normalized to remove the effects of variation in gene copy number, for example owing to variation in plasmid copy number relative to the chromosome or to higher effective copy number near the origin of replication in actively dividing cells. The median for each scaffold was set to zero, and for large scaffolds (over 250 genes), the running median of the gene fitness values was subtracted. Also, for large scaffolds (over 250 genes), the peak of the distribution of gene fitness values was set to zero.

Fitness experiments were deemed successful using the quality metrics that we described previously[9]. These metrics ensure that the typical gene has sufficient coverage, that the fitness values of independent insertions in the same gene are consistent, and that there is no GC bias[9]. Experiments that did not meet these thresholds were excluded from our analyses. The remaining experiments showed good agreement between exact replicates, with a median correlation of 0.87 for gene fitness values from defined media experiments. Ninety-five per cent of the replicate defined media experiments had $r \geq 0.64$. Stress experiments sometimes have little biological signal, as they are usually done in rich medium and mutants of genes that are important for growth in rich medium may be absent from the pools. Nevertheless, the median correlation between replicate stress experiments was 0.74, and 95% of replicate stress experiments had $r \geq 0.35$.

To estimate the reliability of the fitness value for a gene in a specific experiment, we use a *t*-like test statistic, which is the gene's fitness divided by the standard error[9]. The standard error is the maximum of two estimates. The first estimate is based on the consistency of the fitness for the strains in that gene. The second estimate is based on the number of reads for the gene.

Even mild phenotypes were quite consistent between replicate experiments if they were statistically significant. For example, if a gene had a mild but significant phenotype in one replicate ($0.5 < |\text{fitness}| < 2$ and $|t| > 4$), then the sign of the fitness value was the same in the other replicate 96.7% of the time. Because this comparison might be biased if the two replicates were compared to the same control sample, only replicates with independent controls were included.

Genes with statistically significant phenotypes. We averaged fitness values from exact replicate experiments. We combined *t* scores across replicate experiments with two different approaches. If the replicates did not share a start sample and were entirely independent, then we used $t_{\text{comb}} = \text{sum}(t)/\text{sqrt}(n)$, where n is the number of replicates. But if the replicates used the same start sample then this metric would be biased. To correct for this, we assumed that the start and end samples have similar amounts of noise. This is conservative because we usually sequenced the start samples with more than one PCR and with different multiplexing tags. Given this assumption and given that $\text{variance}(A+B) = \text{variance}(A) + \text{variance}(B)$ if A and B are independent random variables, it is easy to show that the above estimate of t_{comb} needs to be divided by $\text{sqrt}((n^2+n)/(2n))$, where n is the number of replicates.

Our standard threshold for a statistically significant phenotype was $|\text{fitness}| > 0.5$ and $|\text{combined } t| > 4$, but this was increased for some bacteria to maintain a false discovery rate (FDR) of less than 5%. We use a minimum threshold on |fitness| as well as |*t*| to account for imperfect normalization or for other small biases in the fitness values. To estimate the number of false positives, we used control experiments, that is, comparisons between different measurements of different aliquots of the same start sample. However we did not use some previously published control comparisons[9] that used the old PCR settings and had strong GC bias (to exclude these, we used the same thresholds that we used to remove biased experiments). The estimated number of false positive genes was then the number of control measurements that exceeded the thresholds, multiplied by the number of conditions and divided by the number of control experiments. As a second approach to estimate the number of false positives, we used the number of expected false positives if the t_{comb} values follow the standard normal distribution ($2 \times \text{P}(z > t) \times$ #experiments $\times$ #genes). If either estimate of the false discovery rate was above 5%, we raised our thresholds for both |fitness| and $|t_{\text{comb}}|$ in steps of 0.1 and 0.5, respectively, until FDR < 5%. The highest thresholds used were $|\text{fitness}| > 0.9$ and $|t| > 6$. Also, for *Pseudomonas fluorescens* FW300-N1B4, we identified six genes with large differences between control samples ($|\text{fitness}| \approx 2$ and $|t| \approx 6$). These genes cluster on the chromosome in two groups and have high cofitness, and several of the genes are annotated as being involved in capsular polysaccharide synthesis. Because this bacterium is rather sticky, we suspect that mutants in these genes are less adherent and were enriched in some control samples owing to insufficient vortexing, so these six genes were excluded when estimating the number of false positives.

Sequence analysis. To assign genes to Pfams[17] or TIGRFAMs[16], we used HMMer 3.1b1[54] and the trusted score cutoff for each family. We used Pfam 28.0 and TIGRFAM 15.0. We used only the curated families in Pfam ('Pfam A').

To identify putative orthologues between pairs of genomes, we used bidirectional best protein BLAST hits with at least 80% alignment coverage both ways. We did not use any cutoff on similarity, as a similarity of phenotype can show that distant homologues have conserved functions. For the analysis of conserved specific phenotypes with cisplatin stress and xylose carbon catabolism, we used greedy clustering to go from pairs of orthologues to orthologue groups. In some instances, a single protein family was split into multiple orthologue groups and combined manually.

To estimate the evolutionary relationships of the bacteria that we studied (Fig. 1b), we used Amphora2[55] to identify 31 highly-conserved proteins in each genome and to align them, we concatenated the 31 protein alignments, and we used FastTree 2.1.8[56] to infer a tree.

Specific phenotypes. We defined a specific phenotype for a gene in an experiment as: $|\text{fitness}| > 1$ and $|t| > 5$ in this experiment; $|\text{fitness}| < 1$ in at least 95% of experiments; and the fitness value in this experiment is noticeably more extreme than most of its other fitness values ($|\text{fitness}| > 95\text{th percentile}(|\text{fitness}|) + 0.5$). Our definition of specific phenotypes is sensitive to the conditions that we profiled. For example, an amino acid biosynthetic gene with an auxotrophic phenotype in defined medium will probably not be identified with a specific phenotype in any of our experiments, as such a gene is expected to have a severe fitness defect in nearly all defined media fitness assays. To minimize this issue for the stress experiments, we chose stress compounds based on the variable genome-wide mutant fitness patterns they elicited from a prior study[3].

To demonstrate the reliability of specific phenotypes on a broader scale, we investigated how often genes with specific phenotypes for the utilization of carbon sources could be assigned to their annotated SEED subsystems[35] (Supplementary Table 21). Across 32 carbon sources, we identified specific-important phenotypes (a specific phenotype with fitness < 0) for 643 genes that are linked to any SEED subsystem, and 388 of these (60%) were linked to the expected subsystem, which is far higher than the 1% that would be expected by chance ($P < 10^{-15}$, Fisher's exact test).

We considered a specific phenotype to be conserved if a potential orthologue had a specific phenotype with the same sign in a similar experiment with the same carbon source, nitrogen source, or stressful compound (but not necessarily using

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

the same base medium or the same concentration of the compound). For specific-important phenotypes (fitness < -1), we also considered a specific phenotype to be conserved if a potential orthologue had fitness < -1 and $t < -4$ in a similar experiment, regardless of how many experiments the orthologue had a phenotype in.

Across our entire data set, many of the specific phenotypes were conserved, which supports their functional relevance. For specific-important phenotypes, if a putative orthologue had a fitness measurement in the corresponding condition, then 34% of the time, the orthologue had fitness < -1. Even for orthologues in the same genus, the phenotype was conserved just 46% of the time, so we believe that the lack of conservation is usually due to indirect phenotypes rather than due to the orthologues having different functions. For strong specific-important phenotypes (fitness < -2) in defined medium conditions, the conservation rate rose to 54% (or 72% within a genus); it appears that strong and specific phenotypes in defined medium are particularly likely to indicate a direct association. In contrast, the conservation rate of specific and detrimental phenotypes (fitness > 1) was just 6%, which suggests that most of these associations are indirect. So, we consider a detrimental association to be conserved only if an orthologue meets the criteria for a specific (and detrimental) phenotype in that condition.

Cofitness and conserved cofitness. We calculate cofitness (r) as the Pearson correlation of all of fitness values for a pair of genes within a single bacterium[3,4]. By contrast, conserved cofitness is identified by comparing cofitness scores for pairs of orthologous genes from two bacteria. Conserved cofitness is defined as minimum(cofitness in the source genome, max(cofitness of orthologues in other genomes)). In other words, a pair of genes is conserved cofit at 0.6 if cofitness > 0.6 in the source genome and any orthologous pair had cofitness > 0.6. Note that the calculations of cofitness and conserved cofitness do not take into account the experimental conditions, so it is common for cofit gene pairs to have different phenotypes in the two bacteria. We also note that the number of cofitness associations in our data are far higher than if the data were just noise. For example, we resampled the gene fitness values from the control comparisons (between 'start' samples) to make a new data set with the same number of experiments in each bacterium but with no biological signal. With these data, there were no cases of conserved cofitness above 0.6 and there were just four genes with cofitness in a bacterium above 0.8.

Our fitness calculations for stress experiments do not correct for the fitness values in the unstressed condition. This leads to the possibility of cofitness between genes that are important for fitness in every condition but might not have any functional relationship. However, even genes that are mildly important for fitness in most conditions often have different phenotypes in a subset of conditions. For example, Extended Data Fig. 1b compares the gene fitness values for *S. loihica* PV-4 growing in LB or in LB with 0.8 mM zinc sulfate. There are 26 genes that have fitness < -2 in both conditions, and as you might expect, these genes are important for fitness in most of the LB stress experiments. So one might expect these genes to give spurious results for cofitness. However, of these 26 genes, 13 are significantly detrimental to fitness in some condition (fitness > 0). Even the remaining 13 genes seem to be important in most, but not all conditions. For example, Shew_1093 (peroxide stress resistance protein *yaaA*) is important for fitness in most conditions but has little or no defect with α-ketoglutarate as the carbon source (two replicates, fitness $= -0.3$ and -0.4). Of the 13 genes that lack a detrimental phenotype, none has high-scoring cofitness ($r > 0.8$) with any gene in *S. loihica* PV-4. Four of these genes did meet our criteria for conserved cofitness, and three of the four cases are biologically plausible (*ruvC* with *recG*, which is also involved in recombinational DNA repair; *recG* with *ruvC*; and *gspL* with other *gsp* genes). More broadly, across 30 bacteria there are 1,797 protein-coding genes in our data set that are important for fitness in our base medium (average fitness < -1). (For the two other bacteria, we do not have fitness data from the base medium.) Of those 1,797 genes, 758 have significant detrimental phenotypes in other conditions, so we do not expect spurious cofitness. Of the remaining 1,039 protein-coding genes, just 343 (33%) met our criteria for a functional association based on cofitness. As illustrated by our analysis of the four such cases in PV-4, we expect that many of these associations are genuine. Therefore, we believe that very few of the thousands of cofitness associations we identified are due to functionally unrelated gene pairs with consistently deleterious phenotypes.

Comparison to other ways of computing cofitness. A related caveat in interpreting the cofitness values is that, if some genes are moderately important for fitness in every condition, they will have high cofitness owing to the varying number of generations in the fitness assays, even if there is no functional relationship. This problem could be avoided if the fitness values were divided by the number of generations for which the cells grew, but we prefer not to rescale the fitness values in this way for several reasons. First, it is not clear what the rescaling should be for motility or survival experiments. Second, we expect that the rescaling would amplify the experimental noise in experiments with fewer generations. Third, any genes whose mutants have strong fitness defects should be absent from our data set anyway, as these mutants should be at very low abundance after the construction of the mutant library and outgrowth from the freezer.

We tested the impact of rescaling the fitness values before computing cofitness for six bacteria (*C. crescentus* N1000, *E. vietnamensis*, *D. japonica* UNC79MFTsu3.2, *K. michiganensis* M5a1, *P. actiniarum*, and *Pedobacter* sp. GW460-11-11-14-LB5). For these bacteria, we have accurate measurements of the total number of generations of growth for every successful fitness assay. Across these six bacteria, if we consider the 2,905 query genes whose most-cofit gene had cofitness above 0.8 (without rescaling), then 95% of these pairs had cofitness above 0.8 after rescaling. Also, in all but three cases, the top cofitness hit was still in the top five hits after we rescaled the cofitness values. The distributions of cofitness values for functionally related gene pairs (same TIGR subrole) or unrelated pairs (different subrole) were also very similar with or without rescaling: for each organism and each set of gene pairs, the Kolmogorov–Smirnov distance (D) between the two distributions was between 0.01 and 0.02. (D is a non-parametric measure of the dissimilarity between two distributions and ranges from 0 for identical distributions to 1 for distributions that do not overlap.) Overall, the cofitness values were very similar regardless of whether the fitness values were scaled by the number of generations or not (Extended Data Fig. 10a).

Also note that we compute cofitness across all experiments, without averaging the replicates. To quantify the impact of averaging the biological replicate fitness experiments on the cofitness values, we computed, for each of the 32 bacteria, the Kolmogorov–Smirnov distance (D) between the distribution of cofitness values for functionally unrelated genes (with different TIGR subroles) as computed with replicates averaged or not. For all 32 bacteria, $D < 0.08$. In other words, the distributions of cofitness are quite similar whether or not the biological replicates are averaged. We also verified that the vast majority of the top cofitness hits were still high-scoring hits if we averaged the replicates before computing cofitness. Of the top cofitness hits with $r \geq 0.8$, 95% were in the top five hits and above 0.8 if we used replicate-averaged cofitness. Another 3% of these cases had replicate-averaged cofitness in the top five hits and between 0.7 and 0.8. Overall, the cofitness values were very similar regardless of whether replicate experiments were averaged or not (Extended Data Fig. 10b). In our data set, 99% of bacterium–condition–concentration combinations have 1–5 replicates, which may explain why averaging the replicates did not have much effect on the cofitness values.

Functional associations and conserved functional associations. In a single bacterium, a gene has a functional association if it has a specific phenotype or high cofitness with another gene. A gene has a conserved functional association if it has a conserved specific phenotype or conserved cofitness.

'Predicting' TIGR subroles from cofitness or conserved cofitness. We considered only cofitness hits that were in the top 10 hits for a query gene, only hits with cofitness above 0.4 (or conserved cofitness above 0.4), and only hits that were at least 20 kb from the query gene. For this analysis, we did not consider nearby genes in the genome because (1) our data have subtle biases from chromosomal position[9], which can inflate cofitness[57]; (2) nearby genes may be in the same operon and may be spuriously cofit because of polar effects (that is, when a transposon insertion in an upstream gene of an operon blocks transcription of the downstream genes); and (3) functional relationships between nearby genes can be identified by comparative genomics[58] and may be less useful. We predict that the gene has the same subrole as the best-scoring hit. When testing cofitness, the hits with the highest cofitness were considered first. When testing conserved cofitness, the hits with the highest conserved cofitness (as defined above) were considered first.

Genome and gene annotations. For previously published genomes, gene annotations were taken from MicrobesOnline, Integrated Microbial Genomes (IMG), or RefSeq. Newly sequenced genomes were annotated with RAST[59], except that *S. koreensis* DSMZ 15582 was annotated by IMG and *P. fluorescens* FW300-N2C3 and *P. fluorescens* FW300-N2E3 were annotated by the NCBI pipeline. See Supplementary Table 22 for a summary of genome annotations and their accession numbers.

Statistical analysis. For Fig. 2b, we chose a sample size of 40 genes with non-conserved phenotypes to limit the manual effort for the analysis. We expected a large difference in the proportions such as 80% versus 40%, and the power to detect such a difference was over 98% for $\alpha = 0.05$.

Classification of how informative annotations are. To assess the existing computational annotations for the 32 genomes, we classified all of their proteins into one of four groups: detailed TIGR role, hypothetical, vague, or other detailed. (1) 'Detailed TIGR role' includes proteins that belong to a TIGRFAM role other than 'Unclassified', 'Unknown function', or 'Hypothetical proteins' and had a subrole other than 'Unknown substrate', 'Two-component systems', 'Role category not yet assigned', 'Other', 'General', 'Enzymes of unknown specificity', 'Domain', or 'Conserved'. (2) 'Hypothetical' includes proteins whose annotation contained the phrase 'hypothetical protein', 'unknown function', 'uncharacterized', or if the entire description matched 'TIGRnnnnn family protein' or 'membrane protein'. (3) Proteins were considered to have 'vague' annotations if the gene description contained 'family', 'domain protein', 'related protein', 'transporter related', or if the entire description matched common non-specific annotations ('abc transporter atp-binding protein', 'abc transporter permease', 'abc transporter substrate-binding

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

protein', 'abc transporter', 'acetyltransferase', 'alpha/beta hydrolase', 'aminohydrolase', 'aminotransferase', 'atpase', 'dehydrogenase', 'dna-binding protein', 'fad-dependent oxidoreductase', 'gcn5-related n-acetyltransferase', 'histidine kinase', 'hydrolase' 'lipoprotein', 'membrane protein', 'methyltransferase', 'mfs transporter', 'oxidoreductase', 'permease', 'porin', 'predicted dna-binding transcriptional regulator', 'predicted membrane protein', 'probable transmembrane protein', 'putative membrane protein', 'response regulator receiver protein', 'rnd transporter', 'sam-dependent methyltransferase', 'sensor histidine kinase', 'serine/threonine protein kinase', 'signal peptide protein', 'signal transduction histidine kinase', 'tonb-dependent receptor', 'transcriptional regulator', 'transcriptional regulators', or 'transporter'). The remaining proteins were considered to have 'other detailed' annotations.

To identify a subset of the proteins annotated as 'hypothetical' or 'vague' that do not belong to any characterized families, we relied on Pfam and TIGRFAM. A Pfam family was considered to be uncharacterized if its name began with either DUF or UPF (which is short for 'uncharacterized protein family'). A TIGRFAM family was considered to be uncharacterized if it had no link to a role or if the top-level role was 'Unknown function'. To identify poorly annotated proteins from diverse bacteria (for Extended Data Fig. 9), we used the rules for vague annotations only.

Protein re-annotation. We manually examined the genes with specific-important phenotypes (fitness < -1 and met criteria for a specific phenotype) on carbon sources in 25 bacteria (we excluded *C. crescentus* NA1000, *D. japonica* UNC79MFTsu3.2, *E. vietnamensis*, *H. seropedicae* SmR1, *K. michiganensis* M5a1, *Pedobacter* sp. GW460-11-11-14-LB5, and *P. actiniarum*). We focused on genes with strongly important phenotypes (fitness < -2) whose annotations (in SEED and/or KEGG) did not imply a role in using that nutrient. In some cases, we also tried to identify the complete catabolic pathway and used cofitness to find additional proteins that were involved but did not meet the threshold for a specific phenotype. For putative transporters, we also considered strong specific-important phenotypes (fitness < -2) on nitrogen sources. The primary tools that we used were the Fitness Browser (which includes Pfam[17], TIGRFAM[16], SEED[35], and the public release of KEGG[34]), MetaCyc[60], PaperBLAST[61], and the Conserved Domains Database (CDD)[62]. (We used SEED/RAST only for the initial annotation of a subset of the genomes in this study, but the Fitness Browser stores the SEED/RAST annotation of every protein sequence.) For proteins that appeared to lack correct annotations in the public release of KEGG, we checked the KEGG website to see whether the annotation had been updated.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data and code availability. The Fitness Browser (http://fit.genomics.lbl.gov) provides access to information from the successful fitness experiments, including details of the experimental conditions, quality metrics for each experiment, per-strain fitness values, gene fitness scores, and *t* values. This Fitness Browser is available at http://fit.genomics.lbl.gov and is archived at https://doi.org/10.6084/m9.figshare.5134840. (This archive includes a SQLite relational database, a BLAST database of protein sequences, and tab-delimited files.) The Fitness Browser also contains the original gene annotations for each of the 32 bacteria we studied; the predicted protein sequences for the annotated genes; the results of various comparative genomics analyses (PFam, TIGRFam, SEED/RAST, and KEGG); and re-annotations based on the fitness data (Supplementary Note 5 and Supplementary Table 12). The Fitness Browser also includes functional associations, but as more fitness experiments are conducted, the functional associations in the Fitness Browser may diverge from the analyses in this manuscript.

Most analyses in this manuscript are available from http://genomics.lbl.gov/supplemental/bigfit (archived at https://doi.org/10.6084/m9.figshare.5134837). The R image contains all of the information in the Fitness Browser except for the per-strain fitness values. The R image also reports: how many insertions were found in each protein-coding gene; the likely-essential genes; metadata and quality metrics for the unsuccessful experiments; gene fitness values from control (start) experiments; which genes had statistically significant phenotypes; the conserved functional associations; the classification of how informative each protein's annotation is; the identifiers and BLAST scores for proteins with hypothetical or vague annotations from diverse bacteria; and the cofitness of pairs of genes with the same TIGR subroles or different TIGR subroles. The website and the archive also include: genome sequences; the mapping between the DNA barcode and the insertion location for each pool of transposon mutants; and which mutant strains were used for computing gene fitness. The analysis of cofitness values based on scaling by the number of generations (or not) is archived separately (https://doi.org/10.6084/m9.figshare.5146309.v1).

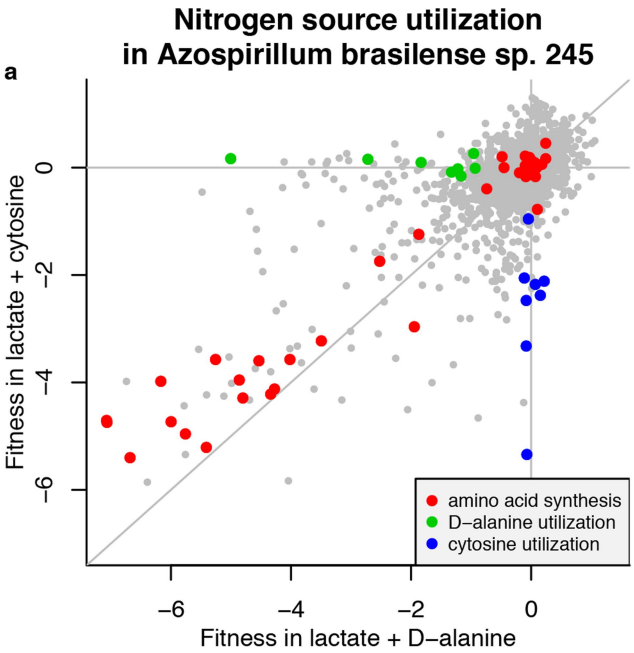
The BarSeq or TnSeq reads were analysed with the RB-TnSeq scripts (https://bitbucket.org/berkeleylab/feba); we used statistics versions 1.0.3, 1.1.0, or 1.1.1 of the code; and version 32×1 of BLAT[63]. The R image was derived from these results using R code that is included in the archive.

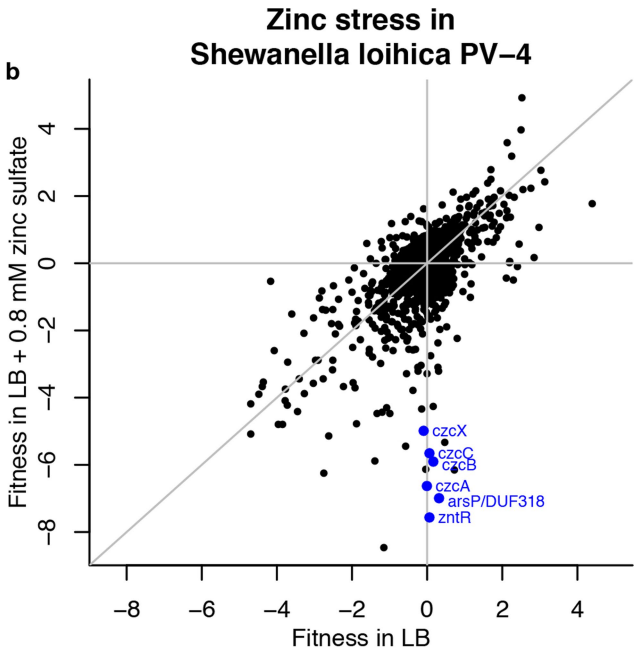
All genomes sequenced for this study have been deposited in GenBank under the following accession numbers: *Acidovorax* sp. GW101-3H11 (LUKZ01000000), *P. fluorescens* FW300-N1B4 (LUKJ01000000), *P. fluorescens* FW300-N2C3 (CP012831.1), *P. fluorescens* FW300-N2E2 (SAMN03294340), *P. fluorescens* FW300-N2E3 (CP012830.1), *P. fluorescens* GW456-L13 (LKBJ00000000.1) and *S. koreensis* DSMZ 15582 (PGEN01000001.1).

Strain availability. Most of the wild-type bacteria (25 of 32), including the eight new isolates from the ORNL FRC, are available from public stock centres. The eight new isolates are available from DSMZ under the following accession numbers: *Acidovorax* sp. GW101-3H11 (DSM 106133), *C. basilensis* 4G11 (DSM 106286), *Pedobacter* sp. GW460-11-11-14-LB5 (DSM 106514), *P. fluorescens* FW300-N1B4 (DSM 106120), *P. fluorescens* FW300-N2C3 (DSM 106121), *P. fluorescens* FW300-N2E2 (DSM 106119), *P. fluorescens* FW300-N2E3 (DSM 106124) and *P. fluorescens* GW456-L13 (DSM 106123). The other seven bacteria were gifts from individuals (Supplementary Table 14). Wild-type bacteria, mutant libraries, and the individual mutant strains that we generated for follow-up studies are available upon request.

40. Thorgersen, M. P. et al. Molybdenum availability is key to nitrate removal in contaminated groundwater environments. *Appl. Environ. Microbiol.* **81**, 4976–4983 (2015).
41. Ray, J. et al. Complete genome sequence of *Cupriavidus basilensis* 4G11, isolated from the Oak Ridge Field Research Center site. *Genome Announc.* **3**, e00322–15 (2015).
42. Vaccaro, B. J. et al. Novel metal cation resistance systems from mutant fitness analysis of denitrifying *Pseudomonas stutzeri*. *Appl. Environ. Microbiol.* **82**, 6046–6056 (2016).
43. Kovach, M. E. et al. Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes. *Gene* **166**, 175–176 (1995).
44. Baba, T. et al. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
45. Kuehl, J. V. et al. Functional genomics with a comprehensive library of transposon mutants for the sulfate-reducing bacterium *Desulfovibrio alaskensis* G20. *MBio* **5**, e01041–14 (2014).
46. Zane, G. M., Yen, H. C. & Wall, J. D. Effect of the deletion of qmoABC and the promoter-distal gene encoding a hypothetical protein on sulfate reduction in *Desulfovibrio vulgaris* Hildenborough. *Appl. Environ. Microbiol.* **76**, 5500–5509 (2010).
47. Kahm, M., Hasenbrink, G., Lichtenberg-Frate, H., Ludwig, J. & Kschischo, M. grofit: fitting biological growth curves with R. *J. Stat. Softw.* **33**, 1–21 (2010).
48. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
49. Bashir, A. et al. A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol.* **30**, 701–707 (2012).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
51. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
52. Tritt, A., Eisen, J. A., Facciotti, M. T. & Darling, A. E. An integrated pipeline for de novo assembly of microbial genomes. *PLoS One* **7**, e42304 (2012).
53. Hunt, M. et al. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294 (2015).
54. Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
55. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).
56. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
57. Sagawa, S., Price, M. N., Deutschbauer, A. M. & Arkin, A. P. Validating regulatory predictions from diverse bacteria with mutant fitness data. *PLoS One* **12**, e0178258 (2017).
58. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA* **96**, 2896–2901 (1999).
59. Aziz, R. K. et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
60. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–D480 (2016).
61. Price, M. N. & Arkin, A. P. PaperBLAST: text mining papers for information about homologs. *mSystems* **2**, e00039–17 (2017).
62. Marchler-Bauer, A. et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222–D226 (2015).
63. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
64. Li, C.-L. et al. DNA binding and cleavage by the periplasmic nuclease Vvn: a novel structure with a known active site. *EMBO J.* **22**, 4014–4025 (2003).
65. Ananthaswamy, H. N. The release of endonuclease I from *Escherichia coli* by a new cold shock procedure. *Biochem. Biophys. Res. Commun.* **76**, 289–298 (1977).
66. Lopes, J., Gottfried, S. & Rothfield, L. Leakage of periplasmic enzymes by mutants of *Escherichia coli* and *Salmonella typhimurium*: isolation of 'periplasmic leaky' mutants. *J. Bacteriol.* **109**, 520–525 (1972).
67. Nossal, N. G. & Heppel, L. A. The release of enzymes by osmotic shock from *Escherichia coli* in exponential phase. *J. Biol. Chem.* **241**, 3055–3062 (1966).

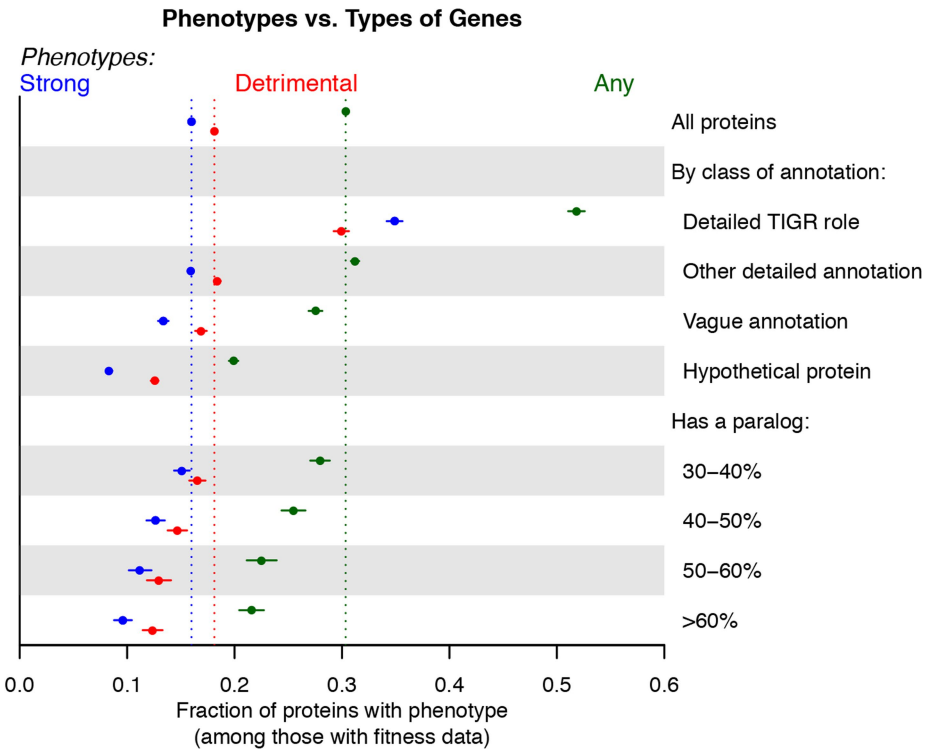
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.





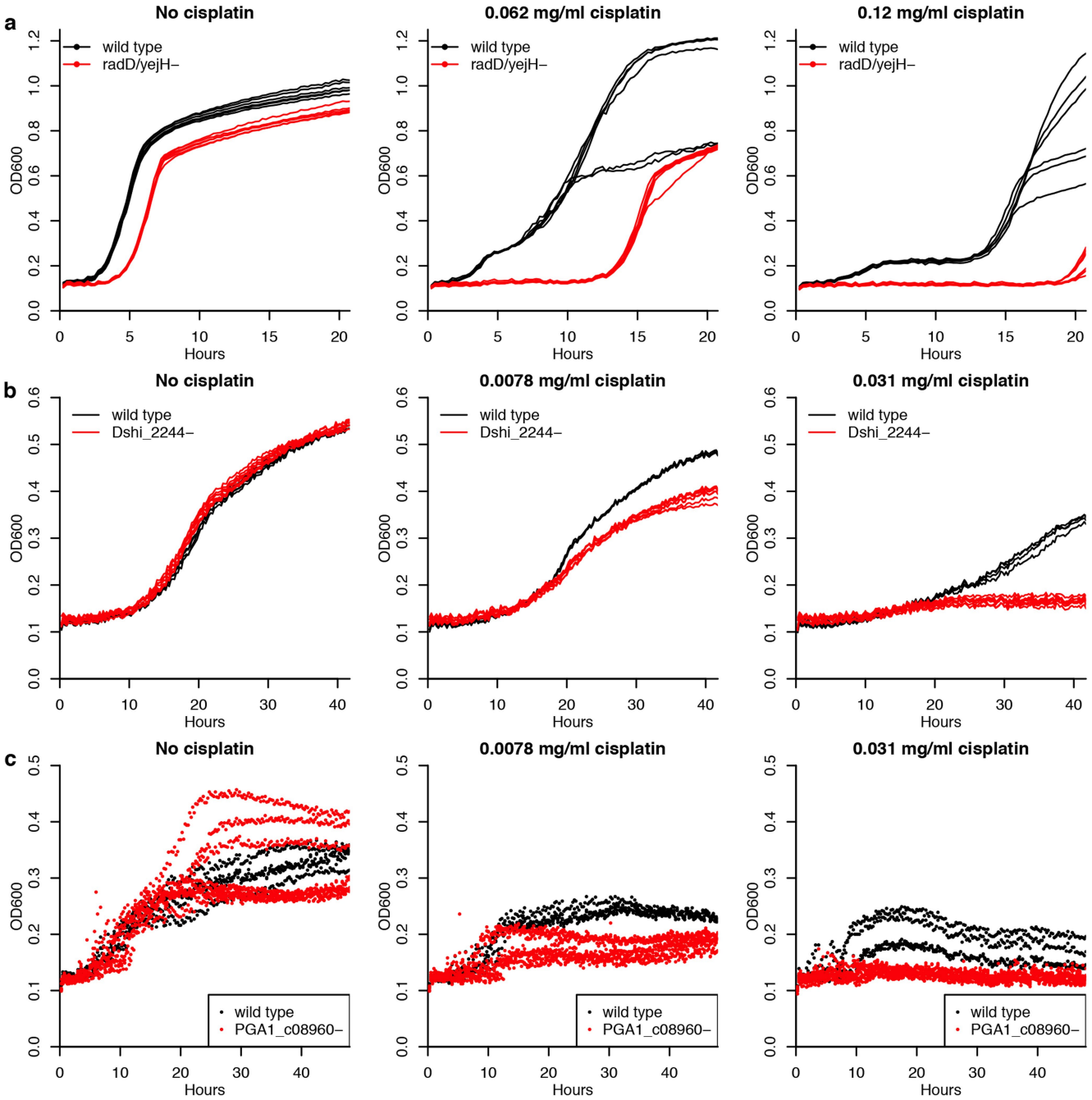
Extended Data Fig. 1 | Examples of nitrogen source and stress fitness experiments. a, The utilization of D-alanine or cytosine by *Azospirillum brasilense* Sp245. Each point shows the fitness of a gene in the two conditions. The data are the average of two biological replicates for each nitrogen source. Amino acid synthesis genes were identified using the top-level role in TIGRFAMs. The genes for D-alanine utilization were a D-amino acid dehydrogenase (AZOBR_RS08020), an ABC transporter operon (AZOBR_RS08235:RS08260), and a LysR family regulator (AZOBR_RS21915). The genes for cytosine utilization were cytosine deaminase (AZOBR_RS31895) and two ABC transporter operons (AZOBR_RS06950:RS06965 and AZOBR_RS31875:RS31885). **b**, Zinc stress in *S. loihica* PV-4. We compare fitness in rich medium with added zinc (II) sulfate to fitness in plain rich medium. The LB data are the average of two biological replicates. The highlighted genes include a putative heavy metal efflux pump (CzcCBA or Shew_3358:Shew_3356), a hypothetical protein at the beginning of the *czc* operon (CzcX), a zinc-responsive regulator (ZntR or Shew_3411), and another heavy metal efflux gene related to *arsP* or DUF318 (Shew_3410). *CzcX* lacks homology to any characterized protein, but homologues in other strains of *Shewanella* are also specifically important for resisting zinc stress. In both panels, the lines show $x = 0$, $y = 0$, and $x = y$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



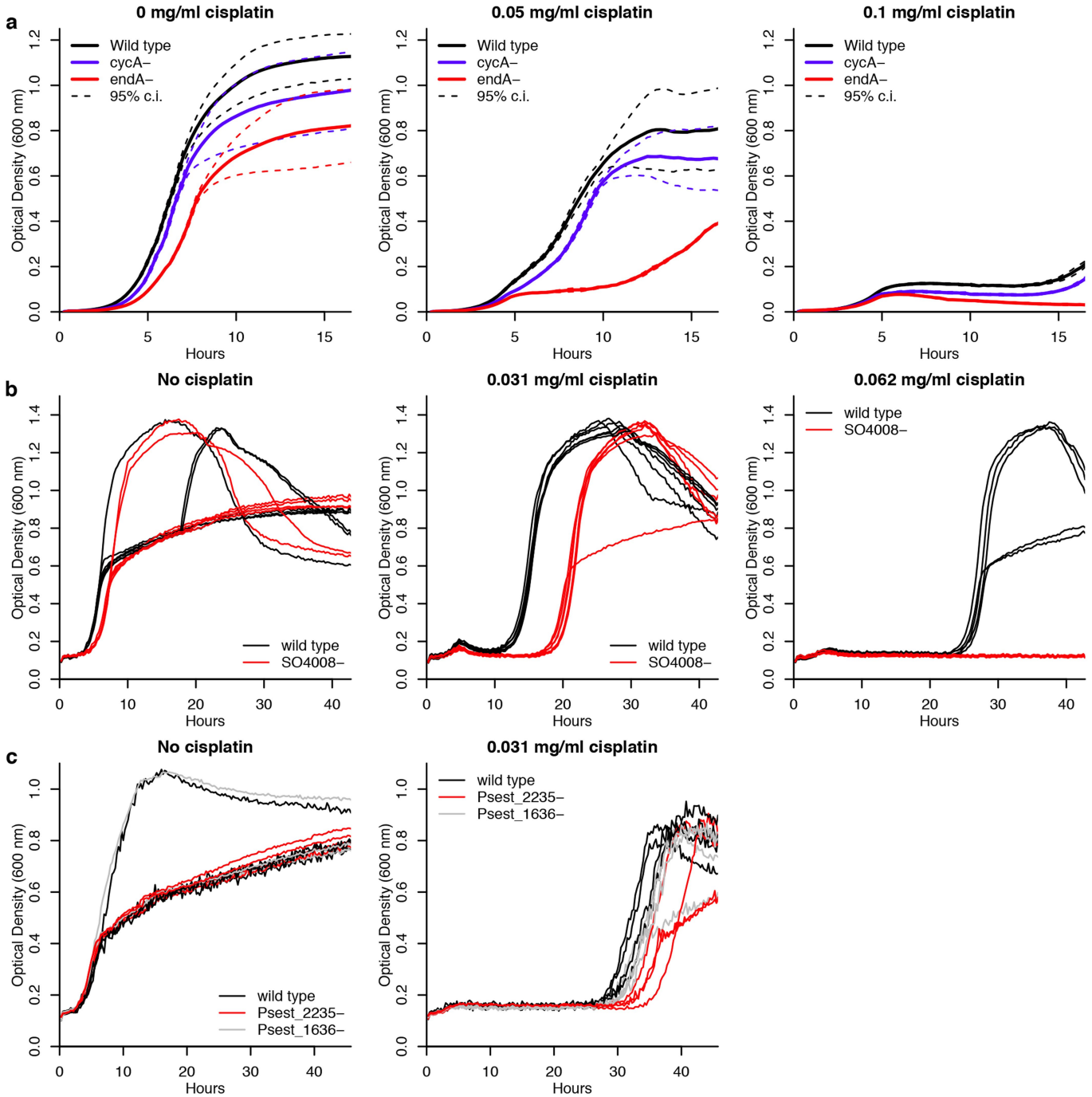
Extended Data Fig. 2 | Phenotypes versus types of genes. We categorized proteins in our data set by their type of annotation or by whether they have homologues in the same genome ('paralogues'). For each category, we show the fraction of genes that have statistically significant phenotypes, and more specifically the fractions that have strong phenotypes ($|\text{fitness}| > 2$ and statistically significant) or are significantly detrimental to fitness (fitness > 0). Genes with high or moderate similarity to another gene in the same genome (paralogues with alignment score above 30% of the self-alignment score) were less likely to have a phenotype (25% versus 32%, $P < 10^{-15}$, Fisher's exact test), which is likely to reflect genetic redundancy.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



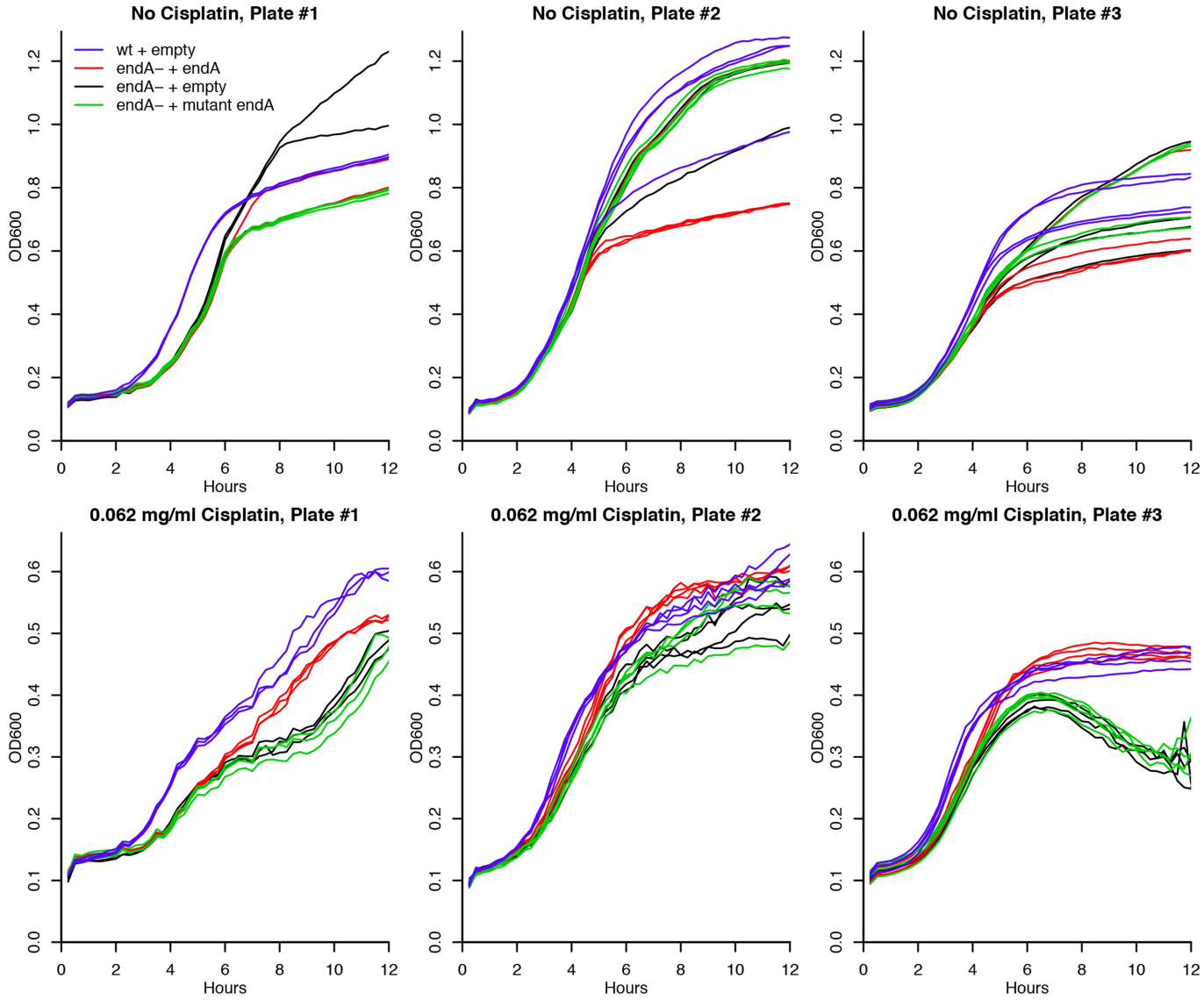
Extended Data Fig. 3 | Known DNA repair genes are important for cisplatin resistance. We compared the growth of a gene deletion strain and the wild-type bacterium under varying cisplatin concentrations. We show all replicate growth curves for each genotype. We believe the higher overall growth for some of the wild-type experiments (for example, top middle) is random. We observe this phenomenon consistently for some bacteria and we speculate that this is due to varying oxygen content across the microplate. **a**, *E. coli radD* ($n = 6$ independent experiments per strain). **b**, *D. shibae* Dshi_2244 ($n = 3$ independent experiments for wild-type and $n = 6$ independent experiments for the mutant). **c**, *Phaeobacter inhibens* PGA1_c08960 ($n = 4$ independent experiments for wild-type and $n = 6$ independent experiments for the mutant). Dshi_2244 and PGA1_c08960 are orthologues of MmcB (DUF1052) from *C. crescentus*[24].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



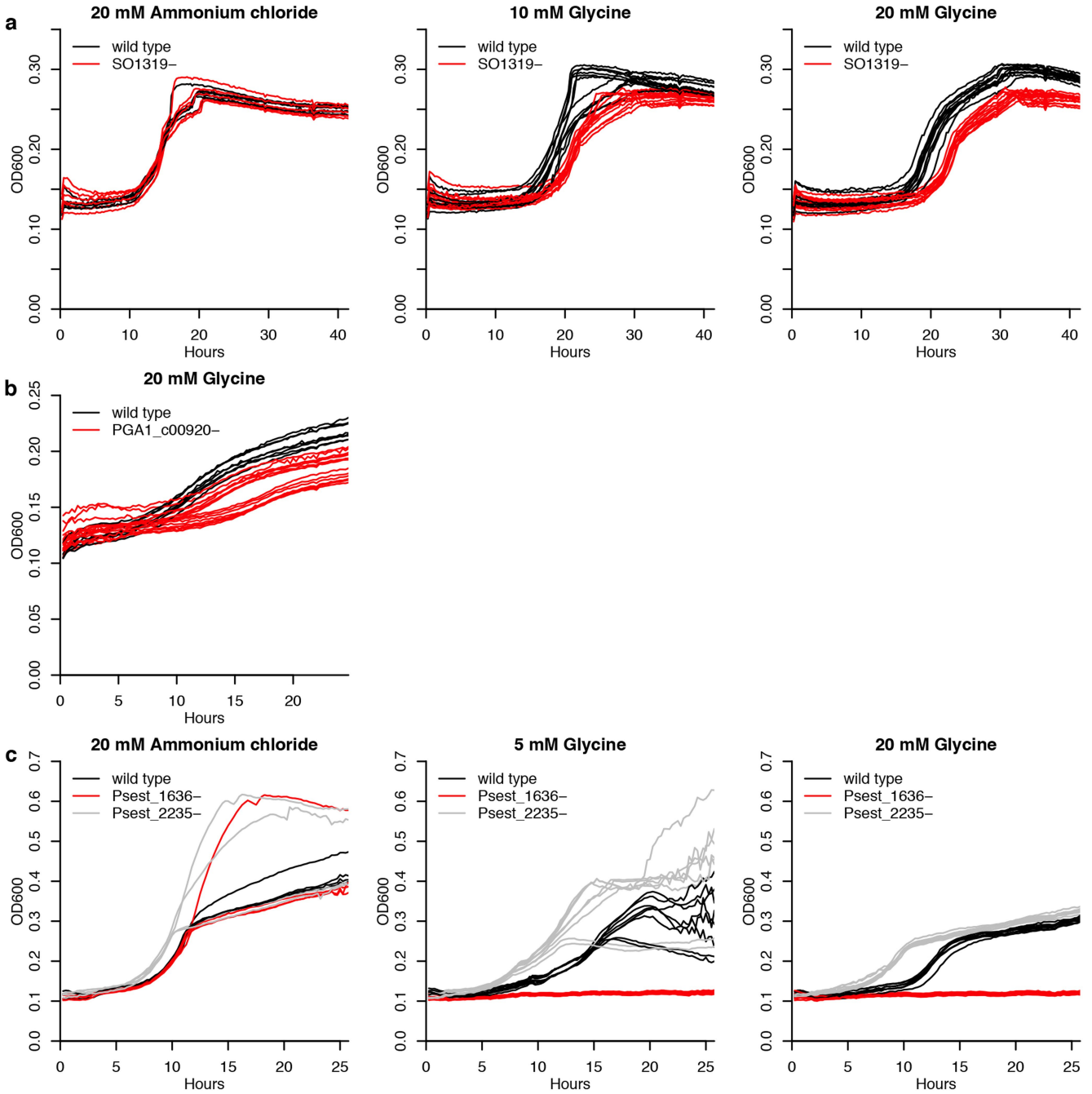
Extended Data Fig. 4 | EndA, DUF3584, and a FAN1-like VRR-NUC domain protein are important for cisplatin resistance. As in Extended Data Fig. 3, comparing cisplatin sensitivity of a gene deletion mutant to the wild-type bacterium. **a**, *E. coli endA* knockout. *cycA* encodes an amino acid transporter and is not expected to have a phenotype on cisplatin and is used as a control. Each growth curve is the average of 12 replicate wells and the dashed lines show 95% confidence intervals from the *t*-test. **b**, A deletion of *S. oneidensis* MR-1 SO4008, a member of the DUF3584 protein family ($n = 6$ independent experiments per strain). **c**, A deletion of *P. stutzeri* RCH2 Psest_2235 ($n = 4$ independent experiments per strain). Psest_1636 is not expected to be involved in DNA repair and is used here as a control. Psest_2235 is a FAN1-like VRR-NUC domain protein[25].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



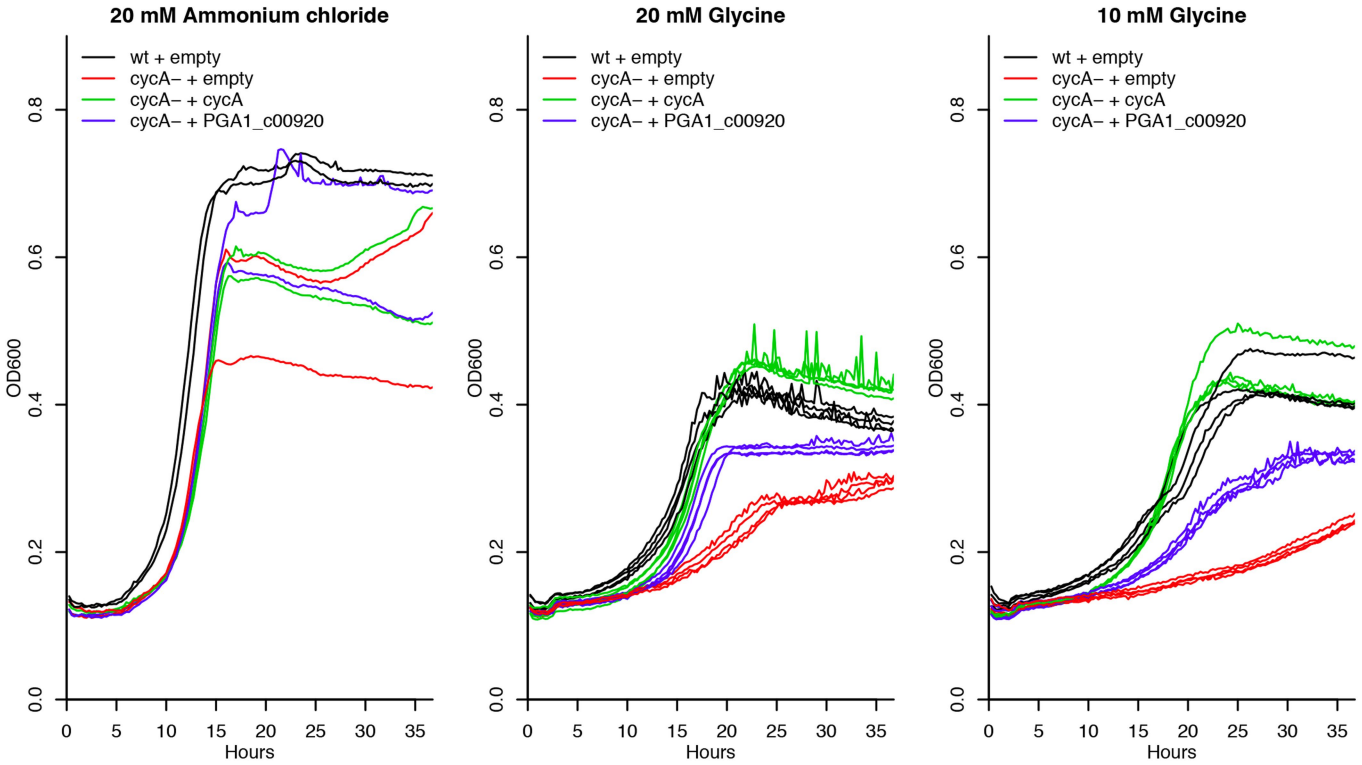
Extended Data Fig. 5 | The nuclease domain of EndA is important for cisplatin resistance. We assayed the growth of an *E. coli endA*− Keio collection deletion mutant carrying one of three different vectors: an empty vector with no insert (endA− + empty), a complementation vector carrying a wild-type copy of *endA* (endA− + endA), and a complementation vector with a mutant version of endA with an alanine at position 84 instead of histidine (endA− + mutant endA). A mutation of this conserved histidine residue in a close homologue from *Vibrio vulnificus* has been reported to eliminate nearly all nuclease catalytic activity[64]. As a control, we assayed the wild-type, parental *E. coli* strain carrying a vector with no insert (wt + empty). We performed these growth assays on three separate microplates (Plate #1, #2, #3). $n = 3$ independent experiments per strain in Plate #1; $n = 4$ independent experiments per strain in Plates #2 and #3. We added 20 μg ml^{-1} gentamicin to each assay to maintain selection for the plasmids (pBBR1-MCS5 and derivatives). Although the catalytic activity of EndA (endonuclease I) appears to be important for resisting cisplatin, it is not clear how EndA would be involved in DNA repair if it is located in the periplasm, as previously believed[65–67]. We speculate that EndA relocates to the cytoplasm upon DNA damage or that EndA degrades broken DNA that enters the periplasm and would otherwise damage the membrane.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



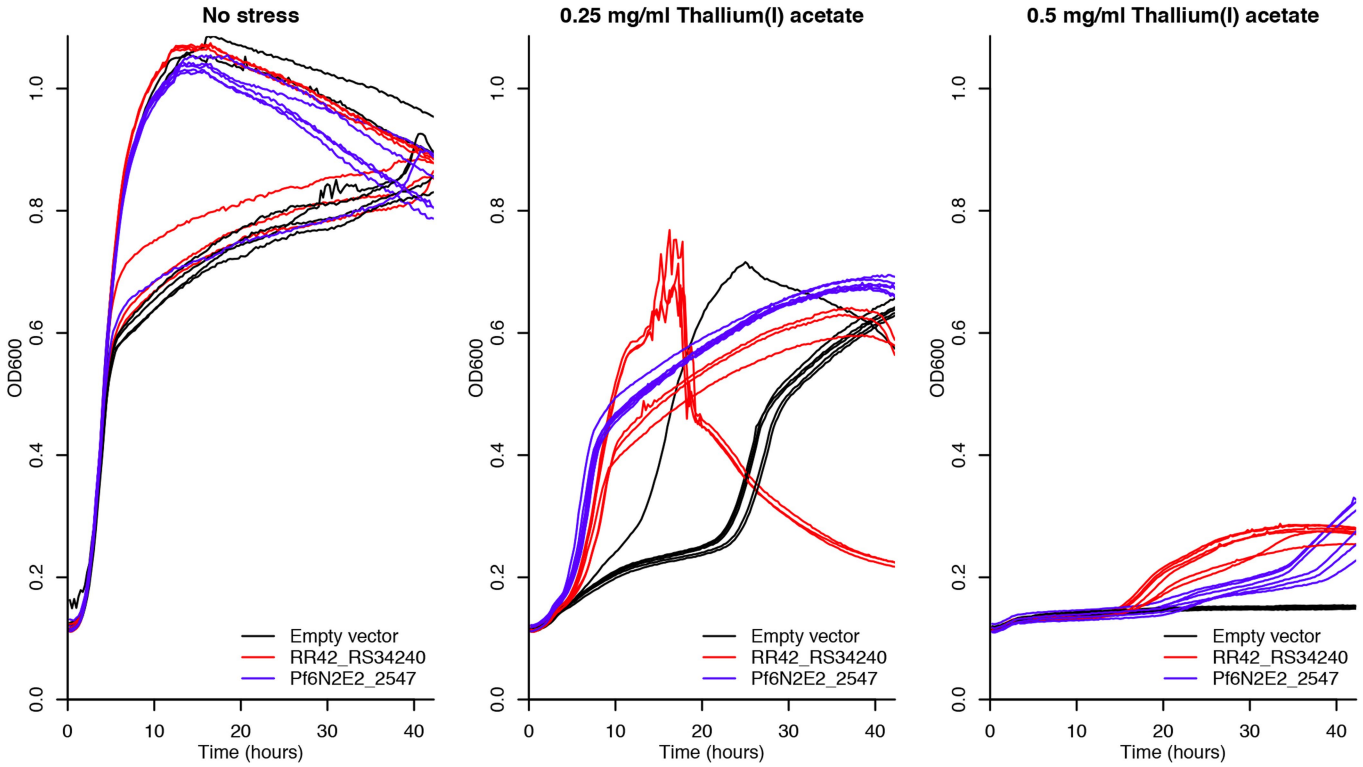
Extended Data Fig. 6 | Members of protein family UPF0126 are important for growth on glycine. Growth comparison of gene deletion mutants in UPF0126 versus wild-type bacteria in minimal defined medium. **a**, SO1319 from *S. oneidensis* MR-1, with either ammonium chloride ($n = 6$ independent experiments per strain) or glycine as the sole source of nitrogen ($n = 12$ independent experiments per strain). **b**, PGA1_c00920 from *P. inhibens*, with glycine as the sole source of carbon ($n = 8$ independent experiments for wild-type and $n = 16$ independent experiments for the mutant). **c**, Psest_1636 from *P. stutzeri* RCH2, with either ammonium chloride ($n = 4$ independent experiments per strain) or glycine ($n = 8$ independent experiments per strain) as the sole source of nitrogen. The Psest_2235 deletion strain is used as a control and is not expected to have a phenotype in these conditions.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



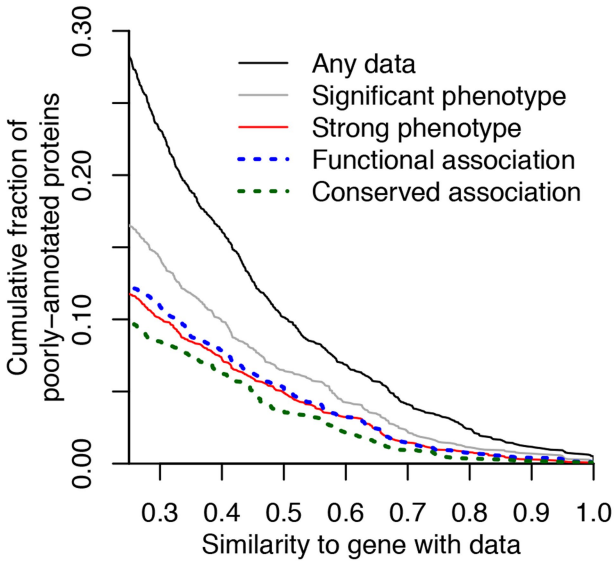
Extended Data Fig. 7 | PGA1_c00920 partially rescues the glycine growth defect of an *E. coli cycA* mutant. CycA is a glycine transporter from *E. coli* and a mutant in this gene has reduced uptake of glycine[37]. We investigated whether a member of the UPF0126 protein family could rescue the glycine growth defect of an *E. coli cycA* deletion strain. We introduced different plasmids into the *E. coli cycA* Keio collection deletion background: an empty plasmid with no insert (cycA− + empty), a plasmid with a wild-type allele of the *E. coli cycA* gene (cycA− + cycA), and a plasmid with PGA1_c00920 from *P. inhibens* (cycA− + PGA1_c00920). We compared the growth of these strains and a wild-type *E. coli* control (wt + empty) in defined media with either ammonium chloride ($n = 2$ independent experiments per strain) or glycine as the sole source of nitrogen ($n = 4$ independent experiments per strain). PGA1_c00920 partially rescues the glycine-specific growth defect of the *cycA*− deletion strain.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



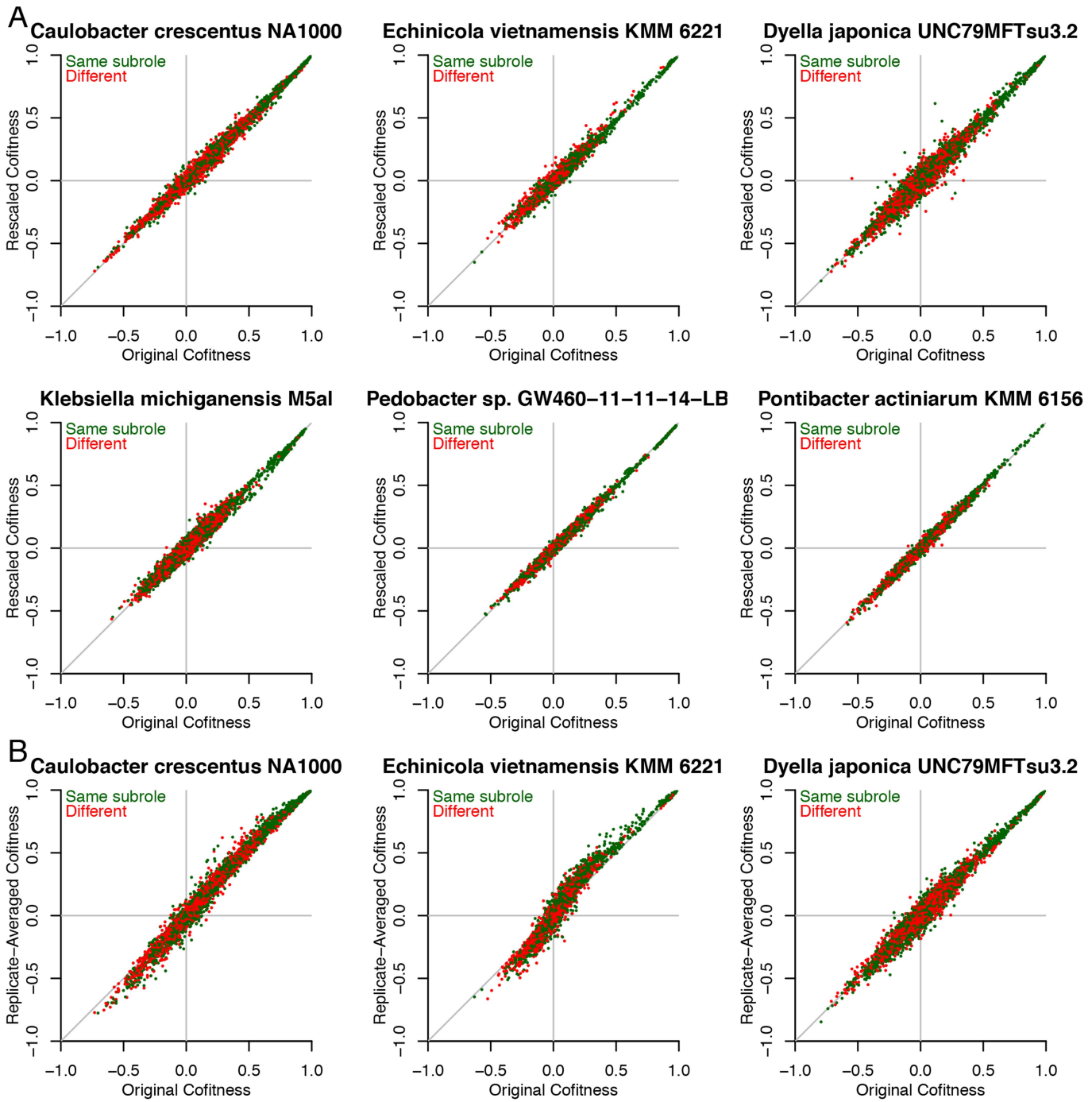
Extended Data Fig. 8 | Overexpression of members of protein family UPF0060 confers resistance to thallium. We introduced three plasmids into wild-type *E. coli*: a plasmid control with no insert (Empty vector), a plasmid carrying RR42_RS34240 from *C. basilensis* 4G11, and a plasmid carrying Pf6N2E2_2547 from *P. fluorescens* FW300-N2E2. We assayed the growth of these strains in LB at 30 °C with varying concentrations of thallium(I) acetate ($n = 6$ independent experiments per strain). We added 50 $\mu g\,ml^{-1}$ kanamycin to each assay to maintain selection for the plasmids (pFAB2286 and derivatives). RR42_RS34240 and Pf6N2E2_2547 are members of the UPF0060 protein family.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | Relevance to all bacteria. We selected 2,593 hypothetical or vaguely annotated proteins from diverse bacterial species, compared them to the protein-coding genes for which we have fitness data (using protein BLAST), and identified potential orthologues as best hits that were homologous over at least 75% of each protein's length. We show the fraction of these proteins that have a potential orthologue with each type of phenotype and that is above a given level of amino acid sequence similarity. Similarity was defined as the ratio of the alignment's bit score to the score from aligning the query to itself.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Alternative ways of computing cofitness. a, The effect of rescaling the cofitness values by the number of generations in six bacteria. For each of the six bacteria, we identified all pairs of protein-coding genes that were assigned to the same TIGR subrole, were more than 20 kB apart, and had fitness data. This gave 1,711–9,406 pairs per bacterium. We also selected a random subset of pairs that were assigned to different TIGR subroles, were more than 20 kB apart, and had fitness data (1,559–8,881 pairs per bacterium). For each pair, we compared the original cofitness values to the rescaled cofitness (computed from fitness values that were divided by the number of generations). **b**, The effect of averaging fitness scores from replicate experiments on the cofitness values.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0137-8

ANKRD16 prevents neuron loss caused by an editing-defective tRNA synthetase

My-Nuong Vo[1,17], Markus Terrey[2,3,4,5,17], Jeong Woong Lee[5,12], Bappaditya Roy[6,7], James J. Moresco[8,13], Litao Sun[1], Hongjun Fu[5,14,15], Qi Liu[6,7,9,16], Thomas G. Weber[10], John R. Yates III[8], Kurt Fredrick[6,7], Paul Schimmel[1,11,18]* & Susan L. Ackerman[2,3,4,5,18]*

Editing domains of aminoacyl tRNA synthetases correct tRNA charging errors to maintain translational fidelity. A mutation in the editing domain of alanyl tRNA synthetase (AlaRS) in *Aars*^{sti} mutant mice results in an increase in the production of serine-mischarged tRNA^{Ala} and the degeneration of cerebellar Purkinje cells. Here, using positional cloning, we identified *Ankrd16*, a gene that acts epistatically with the *Aars*^{sti} mutation to attenuate neurodegeneration. ANKRD16, a vertebrate-specific protein that contains ankyrin repeats, binds directly to the catalytic domain of AlaRS. Serine that is misactivated by AlaRS is captured by the lysine side chains of ANKRD16, which prevents the charging of serine adenylates to tRNA^{Ala} and precludes serine misincorporation in nascent peptides. The deletion of *Ankrd16* in the brains of *Aars*$^{sti/sti}$ mice causes widespread protein aggregation and neuron loss. These results identify an amino-acid-accepting co-regulator of tRNA synthetase editing as a new layer of the machinery that is essential to the prevention of severe pathologies that arise from defects in editing.

Accurate aminoacylation of tRNAs by tRNA synthetases establishes the universal genetic code and occurs in two steps: the activation of the selected amino acid with ATP to form an aminoacyl adenylate, and the transfer of the aminoacyl group of the adenylate to the 2′- or 3′-OH of the cognate tRNA. However, structural similarities between some amino acids allow for misactivation of non-cognate amino acids and subsequent misacylation of tRNA. These errors can be corrected by the hydrolytic editing functions found in many tRNA synthetases[1,2]. Editing can occur after amino acid misactivation but before aminoacyl transfer, or after the aminoacylation of tRNA, known as pre-transfer and post-transfer editing, respectively. Pre-transfer editing occurs at either the active site of aminoacylation or at a distinct editing site, whereas post-transfer editing occurs only at the editing site of these enzymes[3–5].

The importance of tRNA synthetase editing, including editing by the alanyl tRNA synthetase (AlaRS, which is encoded by *Aars* in mice), has been shown in several organisms[6–11]. Although AlaRS misactivates both glycine and serine, the misactivation of serine seems to have more serious consequences, perhaps due to the presence of D-aminoacyl-tRNA deacylases[12–14] that act on Gly-tRNA^{Ala}. Previously we showed that sticky (*sti*) mutant mice that have a point mutation (A734E) in the editing domain of AlaRS have ubiquitinated protein aggregates in cerebellar Purkinje cells, resulting in the degeneration of these neurons[9]. The *sti* mutation results in only a twofold increase in the generation of Ser-tRNA^{Ala}, demonstrating that Purkinje cells are particularly sensitive to the loss of AlaRS editing[9].

Here we identified a new vertebrate-specific gene, *Ankrd16*, which greatly attenuates misfolded aggregate formation and Purkinje cell degeneration in *Aars*$^{sti/sti}$ mice. The deletion of *Ankrd16* in other *Aars*$^{sti/sti}$ neurons also caused the formation of ubiquitinated protein aggregates and neuron death. ANKRD16 binds to AlaRS and prevents the misincorporation of serine at alanine codons in nascent peptides by stimulating serine-dependent ATP hydrolysis before tRNA aminoacylation via the acceptance of misactivated serines. Our data collectively reveal ANKRD16 as a co-regulator of AlaRS that protects against assaults on translation fidelity and proteostasis in mammalian neurons.

Ankrd16 suppresses Purkinje cell loss in sticky mice

Aars$^{sti/sti}$ mice on the inbred C57BL/6J (B6) genetic background are ataxic, with cerebellar Purkinje cell degeneration beginning at three weeks of age[9]. However, neither ataxia nor Purkinje cell degeneration was apparent in nine out of ten 12–16-month-old F_2 *Aars*$^{sti/sti}$ mice from a B6.*Aars*$^{sti/+}$ and CAST/Ei (CAST) mating, suggesting that CAST-derived alleles could suppress neuron loss. In agreement, two phenotypically distinct classes of N_1 *Aars*$^{sti/sti}$ offspring (generated from a backcross of F_1 B6/CAST *Aars*$^{sti/sti}$ mice to B6.*Aars*$^{sti/+}$ mice) were observed at equal frequencies: mice with ataxia and extensive Purkinje cell loss similar to that of B6.*Aars*$^{sti/sti}$ mice, and mice without ataxia and little Purkinje cell loss (Extended Data Fig. 1a). Suppression of the neurodegenerative phenotype was observed in 50% of backcross mice from crosses to CASA/RkJ (CASA, a strain closely related to CAST/Ei), but not in other crosses, which suggests that a single dominant allele from the CAST or CASA strains can suppress *sti*-mediated neurodegeneration (Extended Data Fig. 1a).

Genome scans of affected and unaffected N_1 *Aars*$^{sti/sti}$ mice revealed that the Purkinje cell degeneration was suppressed in mice that were heterozygous for CAST alleles on proximal chromosome 2 (Extended Data Fig. 1b). This locus (around 3.3 Mb) also segregated with the *Aars*$^{sti/sti}$ modifier gene from CASA, which suggests that these strains share the same suppressor gene. Transfer of the modifier of sticky

[1]The Skaggs Institute for Chemical Biology, Department of Molecular Medicine, Scripps Research Institute, La Jolla, CA, USA. [2]Howard Hughes Medical Institute, Department of Cellular and Molecular Medicine, School of Medicine, University of California San Diego, La Jolla, CA, USA. [3]Section of Neurobiology, University of California San Diego, La Jolla, CA, USA. [4]Graduate School of Biomedical Sciences and Engineering, University of Maine, Orono, ME, USA. [5]The Jackson Laboratory, Bar Harbor, ME, USA. [6]Department of Microbiology, The Ohio State University, Columbus, OH, USA. [7]Center for RNA Biology, The Ohio State University, Columbus, OH, USA. [8]Department of Chemical Physiology, Scripps Research Institute, La Jolla, CA, USA. [9]Ohio State Biochemistry Program, The Ohio State University, Columbus, OH, USA. [10]Dynamic Biosensors GmbH, Munich, Germany. [11]The Scripps Research Institute, Jupiter, FL, USA. [12]Present address: Korea Research Institute of Bioscience and Biotechnology, Daejeon, South Korea. [13]Present address: Salk Institute for Biological Studies, La Jolla, CA, USA. [14]Present address: Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY, USA. [15]Present address: Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Medical Center, New York, NY, USA. [16]Present address: Sharklet Technologies, Aurora, CO, USA. [17]These authors contributed equally: My-Nuong Vo, Markus Terrey. [18]These authors jointly supervised this work: Paul Schimmel, Susan L. Ackerman. *e-mail: sackerman@ucsd.edu; schimmel@scripps.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

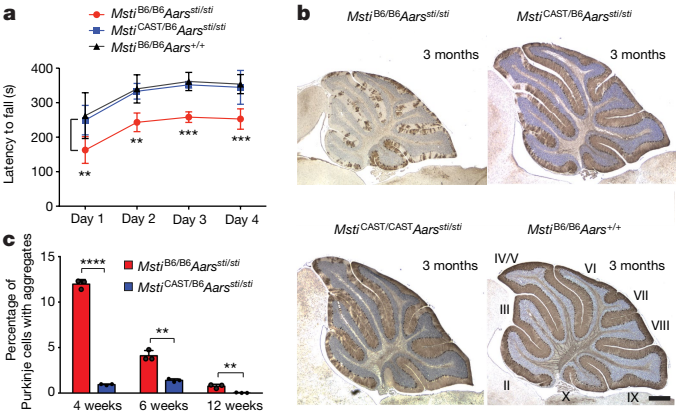


Fig. 1 | Modifier of sticky (*Msti*) suppresses *Aars*^{sti}-mediated neurodegeneration. a, Results of rotarod experiments on three-month-old mice. Data are mean ± s.d., $n = 15$ or 16 mice per genotype, two-way analysis of variance (ANOVA; Tukey correction), $^{**}P \leq 0.01$, $^{***}P \leq 0.001$. **b**, Calbindin D-28 immunohistochemistry images of the indicated genotypes. Sections were counterstained with haematoxylin. $n = 5$ biological replicates. **c**, The percentage of Purkinje cells with aggregates in mice at different ages. Bars and error bars are mean ± s.d., $n = 3$, black circles represent individual data points. Multiple *t*-tests (Holm–Sidak method), $^{**}P = 0.00148$ (6 weeks), $P = 0.00498$ (12 weeks), $^{****}P = 2.819 \times 10^{-6}$. Scale bars, 500 μm.

(*Msti*) locus from the CAST genome onto the B6.*Aars*$^{sti/sti}$ genetic background rescued the decreased latency to fall in a rotarod test of three-month-old B6.*Aars*$^{sti/sti}$ mice and markedly reduced Purkinje cell death, although some neurons in the rostral cerebellum of these mice still degenerated (Fig. 1a, b). Heterozygosity and homozygosity for CAST alleles in the *Msti* region similarly attenuated Purkinje cell loss, in agreement with the dominant nature of *Msti* (Fig. 1b).

Ubiquitin punctae characteristic of protein inclusions were observed in many Purkinje cells in four-week-old B6.*Aars*$^{sti/sti}$ mice, but not in B6 or CAST mice (data not shown). As expected owing to the progressive Purkinje cell loss observed in *sti* mutant mice, the number of Purkinje cells that contained inclusions decreased by 6 and 12 weeks of age. However, very few Purkinje cells in B6.*Msti*$^{CAST/B6}$*Aars*$^{sti/sti}$ mice contained inclusions, even at later ages, which confirms the long-term protection of Purkinje cells (Fig. 1c).

We further localized *Msti* to a 0.63-Mb region, and the coding regions of the protein-coding genes in this interval were amplified by PCR with reverse transcription (RT–PCR) from B6 and CAST cerebellar RNA and sequenced (Fig. 2a). Non-synonymous single nucleotide polymorphisms (SNPs) were observed in *Il2ra*, *Fbxo18* and *Ankrd16* (Fig. 2a). However, with the exception of *Il2ra*, these SNPs were either not present in CASA (*Fbxo18*), which is an *Aars*$^{sti/sti}$ modifying strain, or were also present in MOLF (*Ankrd16*), a strain in which neurodegeneration is not rescued (Extended Data Fig. 1a, c).

A strain-specific *Ankrd16* transcript that contained a 138-bp cryptic exon (exon 5′) was also observed in B6 (with or without the *Aars*$^{sti/sti}$ mutation) and other non-rescuing strains, but was absent in CAST or CASA (Fig. 2b, Extended Data Fig. 1d, e). Sequence analysis of intron 5 suggested that the inclusion of this exon was due to an SNP that generated a novel alternative splice site (Fig. 2c, Extended Data Fig. 1f). Transcript levels of the correctly spliced *Ankrd16* isoform (lacking exon 5′) were 3.3-fold and 5.3-fold higher in the congenic *Msti*$^{CAST/B6}$ or CAST cerebellum relative to B6, respectively (Fig. 2d). Protein levels were 2.8 ± 0.2-fold and 3.9 ± 0.22-fold higher in the B6.*Msti*$^{CAST/B6}$ and CAST cerebellum, respectively, and the *Aars*$^{sti/sti}$ mutation did not alter *Ankrd16* splicing or protein levels (Fig. 2e, Extended Data Fig. 1e, g). Increased ANKRD16 levels were also observed in other tissues from *Msti*$^{CAST/B6}$ mice (Extended Data Fig. 1h). The *Ankrd16* transcript with exon 5′ contains a premature stop codon and is predicted to undergo nonsense-mediated decay; in agreement, truncated forms of ANKRD16 were not observed (Fig. 2e, Extended Data Fig. 1g).

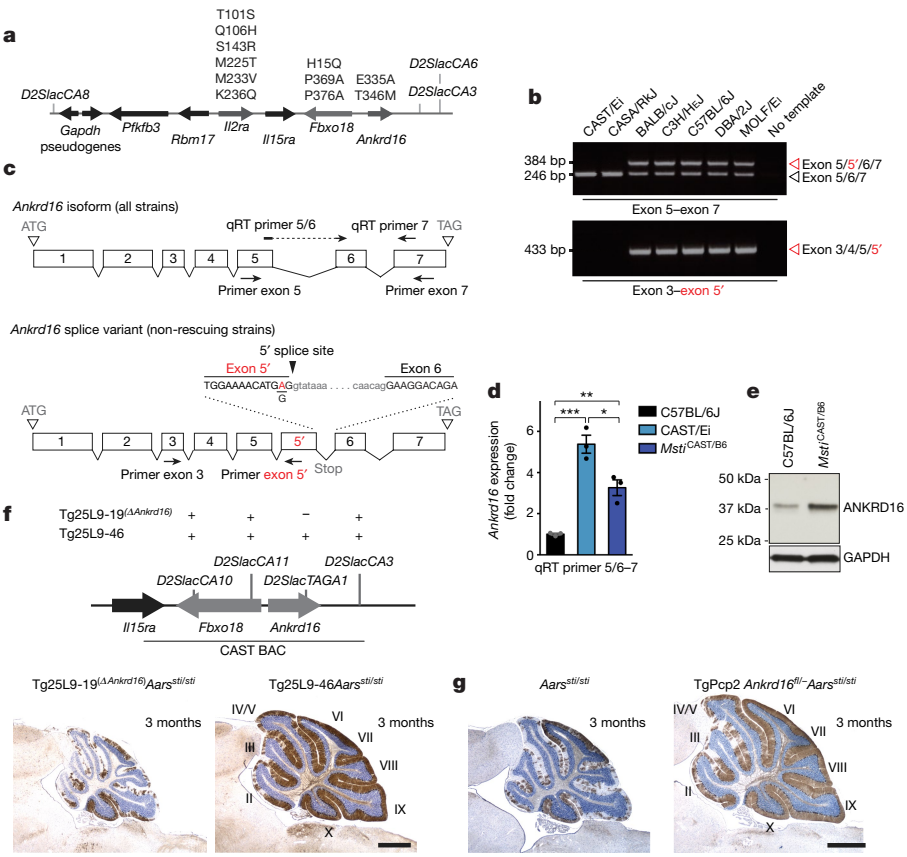


Fig. 2 | *Ankrd16* is the modifier of *Aars*$^{sti/sti}$. **a**, Amino acid polymorphisms between C57BL/6J and CAST/Ei in the *Msti* region. **b**, Alternative splicing of *Ankrd16* in the cerebellum. $n = 2$ biological replicates. **c**, An intronic SNP (red) introduces a cryptic exon (exon 5′) in *Ankrd6* in non-*Msti* containing strains. **d**, Expression of correctly spliced *Ankrd16* in the cerebellum. Bars and error bars are mean ± s.e.m., $n = 3$, black circles represent individual data points. One-way ANOVA (Tukey correction), $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$. **e**, Western blot analysis using cerebellar extracts, $n = 3$ biological replicates. **f**, Schematic of BAC transgene and calbindin D-28 immunohistochemistry images. $n = 5$ biological replicates. **g**, Calbindin D-28 immunohistochemistry images. $n = 5$ biological replicates. Scale bars, 500 μm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

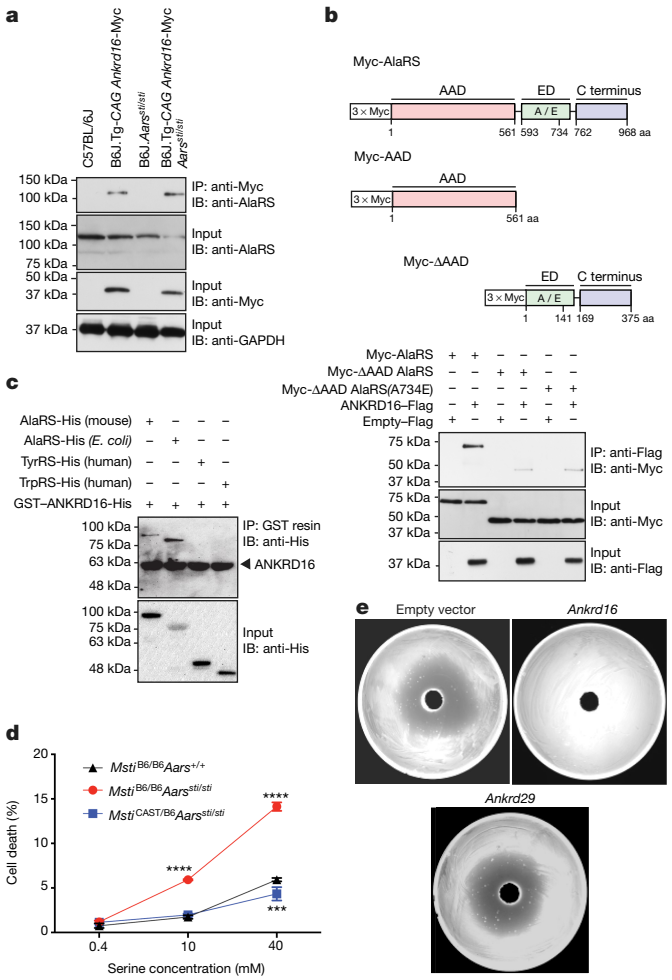


Fig. 3 | ANKRD16 interacts with AlaRS and prevents serine-mediated cell death. a, Co-IP experiments and western blotting reveal the AlaRS–ANKRD16 interaction in the brain. $n = 4$ biological replicates. **b**, Co-IP experiments in HEK293T cells using mouse Myc-AlaRS domain constructs. $n = 2$ independent experiments. **c**, Pull-down of purified AlaRS, TyrRS or TrpRS with ANKRD16. $n = 2$ independent experiments. **d**, Cell death of embryonic fibroblasts. Data are mean ± s.d., $n = 3$, two-way ANOVA (Tukey correction), $^{***}P \leq 0.001$, $^{****}P \leq 0.0001$). **e**, AlaRS(C666A/Q584H) *E. coli* expressing mouse *Ankrd16* or *Ankrd29*. $n = 3$ independent experiments. aa, amino acids; AAD, aminoacylation domain; ED, editing domain; IB, immunoblot; IP, immunoprecipitation.

Neurodegeneration in *Aars*$^{sti/sti}$ mice was modified neither by transgenic expression of the CAST *Il2ra* cDNA nor by deletion of *Il2ra* (data not shown). However, suppression of *Aars*$^{sti/sti}$-mediated Purkinje cell loss was observed in mutant mice that carried a CAST BAC transgene containing the 3′-portion of the *Il15ra* gene, and the *Fbxo18* and *Ankrd16* genes (Fig. 2f). However, another transgenic line (Tg25L9-19) generated with this BAC in which *Ankrd16* was deleted upon integration was not able to suppress Purkinje cell degeneration (Fig. 2f, Extended Data Fig. 1i). Indeed, transgenic expression of the *Ankrd16*^{CAST} coding sequence using the Purkinje cell-specific promoter *Pcp2* in B6.*Aars*$^{sti/sti}$ mice suppressed the death of these neurons (Fig. 2g).

ANKRD16 binds AlaRS to prevent mistranslation

Ankrd16 encodes a 39-kDa protein of unknown function that is composed of nine repeats of the ankyrin protein–protein interaction domain and is found only in vertebrates (Extended Data Fig. 2a, Extended Data Fig. 1j). To identify proteins that interact with ANKRD16, we transgenically expressed *Ankrd16*-Myc under the control of the chicken beta-actin promoter and cytomegalovirus enhancer[15,16]. Co-immunoprecipitation (co-IP) of ANKRD16 was performed using livers of transgenic and non-transgenic mice, and the immunoprecipitates were analysed by liquid chromatography–tandem mass spectrometry (LC–MS/MS). Notably, AlaRS was the most abundant protein after the ANKRD16 bait protein (Extended Data Fig. 2a). Interaction of ANKRD16 with wild-type or mutant AlaRS in both the liver and the brain was confirmed by co-IP and western blotting (data not shown and Fig. 3a). Interactions between ANKRD16 and AlaRS were independent of both the epitope tag and the bait protein (Extended Data Fig. 2b, c).

Co-IP experiments demonstrated that the AlaRS aminoacylation domain was efficiently co-immunoprecipitated by ANKRD16, and little interaction was observed in the absence of this domain (Fig. 3b, Extended Data Fig. 2b). Affinity-capture experiments using purified AlaRS protein further demonstrated that the AlaRS catalytic domain was sufficient for interaction with ANKRD16 and that this interaction was direct (Extended Data Fig. 2e, Fig. 3c). Furthermore, AlaRS did not co-immunoprecipitate ANKRD29—a protein that comprises eight ankyrin repeats (Extended Data Fig. 2d)—nor did ANKRD16 interact with tyrosyl-tRNA synthetase or tryptophanyl-tRNA synthetase (Fig. 3c). This suggests a specific ANKRD16–AlaRS interaction, with similar binding dynamics for both wild-type AlaRS and AlaRS(A734E) (Extended Data Fig. 2f).

The AlaRS(A734E) mutation leads to increased death of *Aars*$^{sti/sti}$ embryonic fibroblasts when cultured with increasing concentrations of serine[9]. *Msti*$^{CAST/B6}$*Aars*$^{sti/sti}$ fibroblasts were more resistant to high serine concentrations than were B6.*Aars*$^{sti/sti}$ fibroblasts; at the highest concentrations (40 mM), less cell death was observed in *Msti*$^{CAST/B6}$*Aars*$^{sti/sti}$ than in B6 (*Msti*$^{B6/B6}$) fibroblasts, which demonstrates that *Ankrd16* suppresses serine-mediated cell death in *Aars*$^{sti/sti}$ cells (Fig. 3d).

Although ANKRD16 is a vertebrate-specific protein , mammalian and *Escherichia coli* AlaRS are highly conserved with around 67% similarity in the catalytic domain. Indeed, mouse ANKRD16 was able to bind *E. coli* AlaRS in affinity-capture experiments (Fig. 3c), and reduced the death of *E. coli* with a severe editing-domain mutation (C666A/Q584H) in AlaRS when grown on a serine gradient (Fig. 3e). Together, these results suggest that ANKRD16 influences the function of AlaRS.

Next, we used dipeptide formation on defined ribosome complexes as a read-out of charging and decoding fidelity. These experiments sufficiently detected serine mistranslation and AlaRS-mediated editing of Ser-tRNA^{Ala} (Extended Data Fig. 3a, b). When reactions using Ser-tRNA^{Ala} and AlaRS were supplemented with alanine, a loss of fMet-Ser (fMet, *N*-formylmethionine) coincided with a gain of fMet-Ala (Fig. 4a), which is consistent with the reported post-transfer editing by AlaRS[7]. However, no obvious differences were observed between reactions using wild-type AlaRS or AlaRS(A734E), and ANKRD16 had no effect on either reaction in these experiments using mischarged tRNA.

By contrast, when deacyl tRNA^{Ala}, serine and ATP were used, AlaRS(A734E) generated more fMet-Ser than did wild-type AlaRS, and ANKRD16 reduced the amount of this aberrant dipeptide in reactions with either AlaRS(A734E) or wild-type AlaRS (Extended Data Fig. 3c). Although ANKRD16 interacts with a truncated AlaRS (AlaRS(1–455)) (Extended Data Fig. 2f), no reduction in the amount of aberrant dipeptide was observed, which suggests that the editing and/or C-terminal domains of AlaRS may contribute to the function of ANKRD16. Furthermore, in the presence of ATP and a mixture of serine and alanine, ANKRD16 specifically decreased the formation of fMet-Ser, thereby increasing AlaRS(A734E) fidelity (that is, fMet-Ala/fMet-Ser) by a factor of around 20 (Fig. 4b). Therefore, ANKRD16 acts before the formation of mischarged tRNA to enhance translational fidelity.

In agreement, ANKRD16 had no effect on subsequent steps of tRNA mischarging, including deacylation of mischarged Ser-tRNA^{Ala}, transfer of the mischarged tRNA to EF-Tu, or amino acid activation or aminoacylation of the cognate amino acid alanine (Extended Data Fig. 3d–g). In addition to post-transfer editing activity, pre-transfer editing has also been observed in AlaRS in *E. coli*[17]. Serine-dependent ATPase activity was also observed with mouse wild-type AlaRS in

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

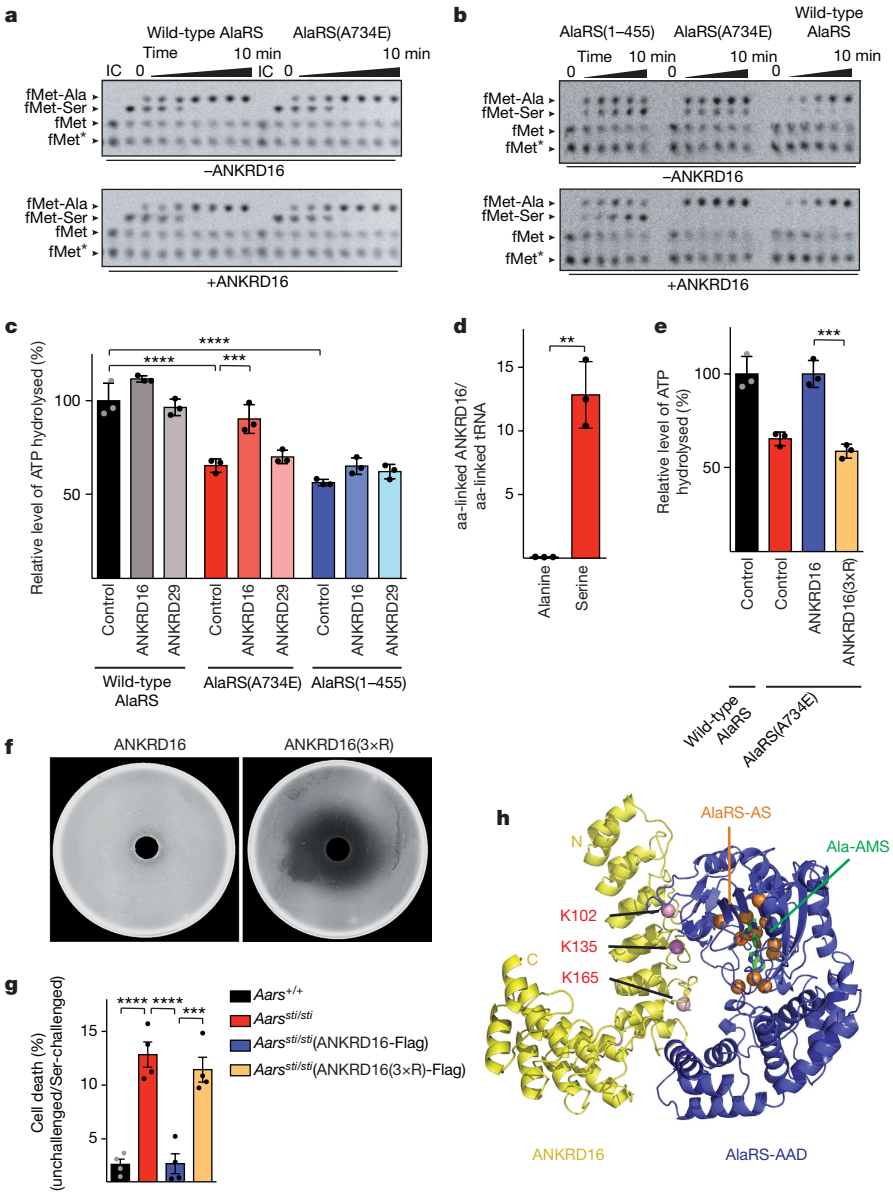


Fig. 4 | ANKRD16 enhances tRNA-independent editing and accepts serine adenylates from AlaRS. a, Mouse AlaRS (±ANKRD16) was added to reactions containing Ser-tRNA^{Ala} and alanine. $n = 2$ independent experiments. **b**, Mouse AlaRS (±ANKRD16) was added to reactions containing deacylated tRNA^{Ala}, alanine and serine (* indicates oxidized fMet). $n = 2$ independent experiments. **c**, Ser-AMP hydrolysis at 10 min. Bars and error bars are mean ± s.d., $n = 3$, black circles represent individual data points. One-way ANOVA (Tukey correction), $^{***}P \leq 0.001$, $^{****}P \leq 0.0001$. **d**, Ratio of alanine- or serine-linked ANKRD16 to aminoacylated tRNA^{Ala}. Bars and error bars are mean ± s.d., $n = 3$, black circles represent individual data points. Two-tailed Student's *t*-test, $^{**}P = 0.0011$. **e**, Ser-AMP hydrolysis at 10 min. Bars and error bars are mean ± s.d., $n = 3$, black circles represent individual data points. One-way ANOVA (Tukey correction), $^{***}P \leq 0.001$. Control reactions are from **c**. **f**, AlaRS(C666A/Q584H) *E. coli* expressing mouse ANKRD16 or ANKRD16(3×R). $n = 2$ independent experiments. **g**, Cell death of embryonic fibroblasts. Bars and error bars are mean ± s.e.m., $n = 4$, circles represent individual data points. One-way ANOVA (Tukey correction), $^{***}P \leq 0.001$, $^{****}P \leq 0.0001$. **h**, Structural model of ANKRD16 and the aminoacylation domain (AAD) of AlaRS. Modified lysines are shown in purple, active site (AS) residues in orange, and Ala-AMS (Ala-AMP analogue) in green.

the absence of tRNA (Extended Data Fig. 4a). However, this activity was decreased with both AlaRS(A734E) and AlaRS(1–455) (Fig. 4c, Extended Data Fig. 4b). ANKRD16 was able to restore that activity for AlaRS(A734E), but not for AlaRS(1–455), which further suggests that the editing domain contributes to the serine-dependent ATPase activity and/or the modulation of this activity by ANKRD16 (Fig. 4c, Extended Data Fig. 4b, c).

Unexpectedly, ANKRD16 appeared to stimulate the tRNA-dependent precipitation of [^{3}H]serine but not of [^{3}H]alanine in reactions with AlaRS(A734E) (Extended Data Fig. 4d–f). However, misactivated [^{3}H]serine was also precipitated in reactions without tRNA^{Ala}, with approximately three linked-serine adducts per ANKRD16 (Extended Data Fig. 4g). By contrast, [^{3}H]alanine-adduct formation per ANKRD16 with AlaRS(A734E), in the absence of tRNA^{Ala}, was much lower (Extended Data Fig. 4h). Furthermore, in the presence of tRNA^{Ala} and AlaRS(A734E), alanine linked to ANKRD16 was negligible relative to that coupled to tRNA^{Ala} (Fig. 4d).

These results suggest that ANKRD16 functions as an alternative to water or tRNA^{Ala}, in that it accepts misactivated serine. Previous studies have demonstrated that reactive aminoacyl adenylates may react with cysteine or lysine residues[18–21]. Levels of precipitated serine were diminished only slightly after prolonged exposure to alkaline pH, which hydrolyses thiol esters (for example, cysteine). As such, the ANKRD16 serine adduct is probably an amide (for example, lysine) (Extended Data Fig. 4i).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

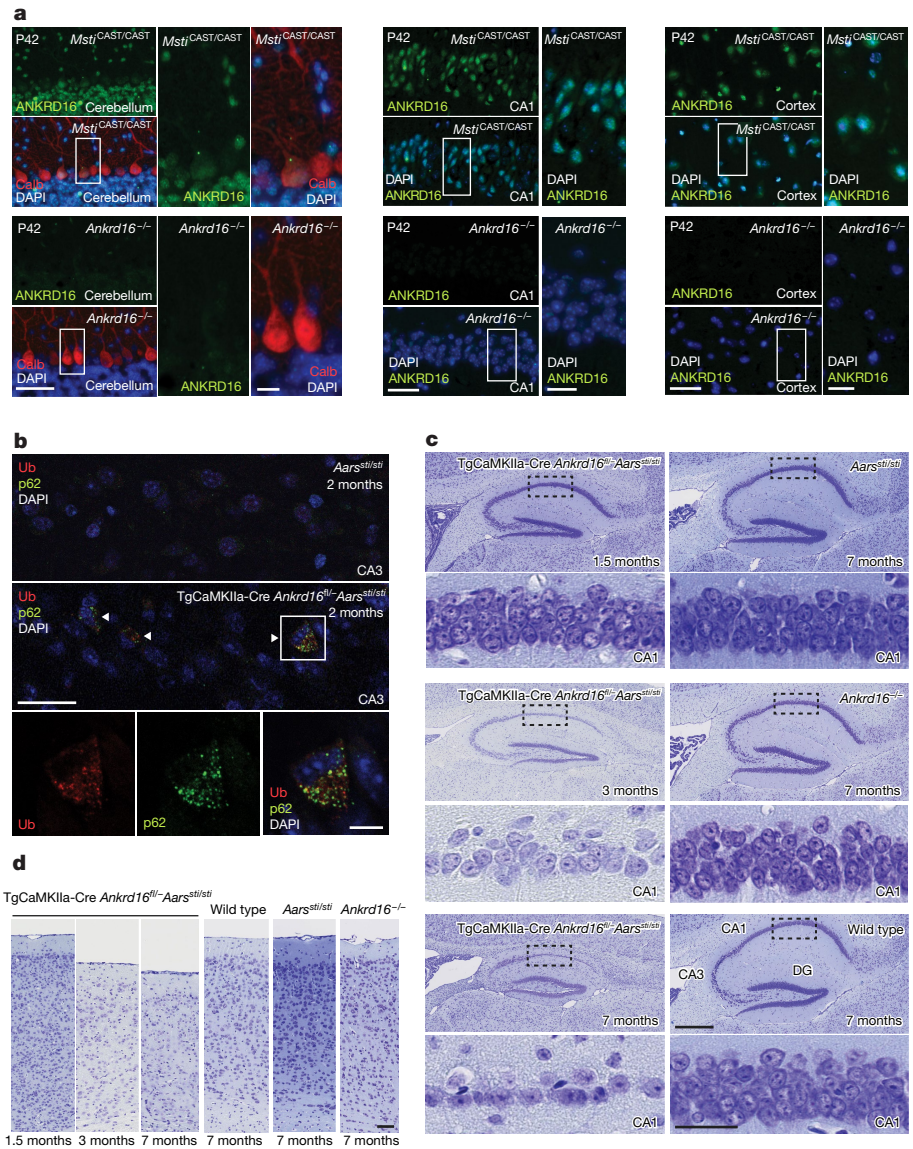


Fig. 5 | Loss of *Ankrd16* in *Aars*$^{sti/sti}$ mice causes protein aggregation and neurodegeneration. **a**, **b**, Immunofluorescence with antibodies to ANKRD16 (green) and calbindin D-28 (Calb; red) (**a**), or ubiquitin (Ub; red) and p62 (green) (**b**). $n = 3$ biological replicates. **c**, **d**, Cresyl violet staining of hippocampus (**c**) and cortex (**d**). Biological replicates for **c**, **d**: TgCaMKIIa-Cre *Ankrd16*$^{fl/-}$*Aars*$^{sti/sti}$: 1.5 months, $n = 5$; 3 months, $n = 7$; 7 months, $n = 6$; *Ankrd16*$^{-/-}$: 7 months, $n = 7$; *Aars*$^{sti/sti}$ and wild-type: 7 months, $n = 3$. Scale bars, 50 μm (**a**, **b**, low magnification; **c**, higher magnification); 10 μm (**a**, **b**, higher magnification); 500 μm (**c**, low magnification); 100 μm (**d**). CA, cornu ammonis; DAPI, 4′,6-diamidino-2-phenylindole; DG, dentate gyrus; P, postnatal day.

We used mass spectrometry to identify the residues of ANKRD16 that accept misactivated serine. No spectral shift was observed with Cys-containing peptides. Although not all Lys-containing peptides of ANKRD16 were resolved, three highly conserved lysines—K102, K135 and K165—shifted by 87 Da, the mass of serine (Extended Data Fig. 1j, Extended Data Fig. 5a–d). Therefore, we mutated those lysine codons to arginine (ANKRD16(K102R/K135R/K165R, hereafter denoted ANKRD16(3×R)) (Extended Data Fig. 4g). Circular dichroism spectroscopy and thermal shift assays suggested that wild-type ANKRD16 and ANKRD16(3×R) have similar secondary structures and stabilities (Extended Data Fig. 5e, f), and only a slight change (approximately twofold) in affinity for AlaRS(A734E) (Extended Data Figs. 2f, 5g). In contrast to ANKRD16, ANKRD16(3×R) had no effect on AlaRS(A734E) pre-transfer editing (Fig. 4e, Extended Data Fig. 4b). Unlike ANKRD16, which rescued the growth of AlaRS(C666A/Q584H) *E. coli* at high concentrations of serine, ANKRD16(3×R) did not rescue the growth of mutant bacteria under these conditions (Fig. 4f). Similarly, the expression of ANKRD16, but not ANKRD16(3×R), prevented the serine-induced death of B6.*Aars*$^{sti/sti}$ embryonic fibroblasts (Fig. 4g, Extended Data Fig. 6). Accordingly, we generated a model for the ANKRD16–AlaRS complex (Fig. 4h). Modified lysine side chains of ANKRD16 project out from the helix–loop–helix motifs, with each in close proximity to the active site of AlaRS, at which serine is misactivated.

Extensive neuron loss in sticky mice lacking *Ankrd16*

We performed immunofluorescence experiments to determine whether the specificity of neuron death in *Aars*$^{sti/sti}$ mice is correlated with the levels of ANKRD16. ANKRD16 was widely expressed in the brain and was detected in the nucleus as well as the cytoplasm (Fig. 5a, Extended Data Fig. 7a, b). Notably, ANKRD16 levels were lower in Purkinje cells relative to cerebellar granule cells, cells of the cerebellar molecular layer, hippocampal neurons, and cells in the cortex (Fig. 5a).

To test whether the levels of *Ankrd16* influence the sensitivity of cells to the *Aars*$^{sti/sti}$ mutation, we generated an *Ankrd16*-null allele (Extended Data Fig. 7c, d). *Ankrd16*$^{-/-}$ mutant mice had no obvious pathologies in the brain or other organs, even when aged to 12 months (data not shown). However, loss of *Ankrd16* in *Aars*$^{sti/sti}$ mice resulted in early embryonic lethality (Extended Data Fig. 7e). These results show that the low levels of ANKRD16 present in B6 mice are sufficient to resolve *sti*-mediated editing defects induced during embryonic development.

To determine the effect of decreasing ANKRD16 levels on neuronal cell survival, we conditionally deleted *Ankrd16* in postnatal Purkinje cells in *Aars*$^{sti/sti}$ mice. Purkinje cell loss in TgPcp2-Cre *Ankrd16*$^{fl/-}$*Aars*$^{sti/sti}$ mice began at about three weeks of age; by four weeks of age, the majority of Purkinje cells were absent and, in contrast to *Aars*$^{sti/sti}$ mice, all Purkinje cells had degenerated by seven months of age (Extended Data Figs. 8a, 9a). Formation of ubiquitin- and p62-positive aggregates was

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

also accelerated in Purkinje cells, with aggregates observed in 12.3% of Purkinje cells compared to 2.5% of these neurons in *Aars*$^{sti/sti}$ mice at three weeks of age (Extended Data Fig. 9a).

Deletion of *Ankrd16* in the embryonic cerebellar primordium using *En1*^{cre}[22] resulted in protein aggregates and neuron death in interneurons in the molecular layer and neurons in the granule cell layer as well as in the Purkinje cells (Extended Data Figs. 8b, c, 9b). Furthermore, loss of *Ankrd16* in postnatal cortical and hippocampal neurons also caused their degeneration. Ubiquitin- and p62-positive aggregates were observed both in hippocampal pyramidal cells ($2.80 \pm 0.22\%$) and in cortical neurons in two-month-old TgCaMKIIa-Cre *Ankrd16*$^{fl/-}$*Aars*$^{sti/sti}$ mice (Fig. 5b–d, Extended Data Fig. 9c). Together, these data demonstrate that neurons other than Purkinje cells are also sensitive to the effects of mistranslation, and further suggest that ANKRD16 protects against mistranslation in a dose-dependent fashion.

Discussion

In contrast to known editing mechanisms of tRNA synthetases or free-standing homologues of editing domains that act autonomously[23,24], tRNA-independent hydrolysis of misactivated serine by AlaRS is enhanced through the binding of ANKRD16, which captures misactivated serine and removes it from the pool for protein synthesis (Extended Data Fig. 10a, b). Unlike most tRNA synthetases, AlaRS misactivates amino acids that are both smaller (glycine) and larger (serine) than alanine, owing to structural properties that make it difficult for AlaRS to exclude misactivated serine[2,24–27]. Thus, serine misactivation by AlaRS, and the toxic effects of serine-for-alanine replacements in vertebrates, may present a special situation in which an editing co-factor is necessary for proofreading.

In addition to providing direct mechanistic insights into the editing functions of aminoacyl tRNA synthetases, the discovery of *Ankrd16* highlights the importance of studying mRNA translation in higher organisms, and may provide understanding for the cell-type sensitivity of phenotypes associated with the *Aars*^{sti} mutation. More broadly, cell-type selectivity resulting from 'monogenic' mutations in ubiquitous genes has been difficult to resolve. Only a few modifier genes of disease mutations have been identified in an unbiased approach, and these suggest that modifier genes may function in independent, parallel pathways[28–30]. Our identification of *Ankrd16* as a modifier of the *Aars*^{sti} mutation demonstrates that this possibility may be an oversimplification, and that restricted pathologies may be due to the expression levels of genes that modify the function of the gene harbouring the primary mutation.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0137-8.

Received: 17 April 2017; Accepted: 9 April 2018;
Published online: 16 May 2018

1. Guo, M. et al. Paradox of mistranslation of serine for alanine caused by AlaRS recognition dilemma. *Nature* **462**, 808–812 (2009).
2. Schimmel, P. Mistranslation and its control by tRNA synthetases. *Philos. Trans. R. Soc. Lond. B* **366**, 2965–2971 (2011).
3. Ling, J., Reynolds, N. & Ibba, M. Aminoacyl-tRNA synthesis and translational quality control. *Annu. Rev. Microbiol.* **63**, 61–78 (2009).
4. Martinis, S. A. & Boniecki, M. T. The balance between pre- and post-transfer editing in tRNA synthetases. *FEBS Lett.* **584**, 455–459 (2010).
5. Dulic, M., Cvetesic, N., Perona, J. J. & Gruic-Sovulj, I. Partitioning of tRNA-dependent editing between pre- and post-transfer pathways in class I aminoacyl-tRNA synthetases. *J. Biol. Chem.* **285**, 23799–23809 (2010).
6. Schimmel, P. Development of tRNA synthetases and connection to genetic code and disease. *Protein Sci.* **17**, 1643–1652 (2008).
7. Beebe, K., Pouplana, L. R. De & Schimmel, P. Elucidation of tRNA-dependent editing by a class II tRNA synthetase and significance for cell viability. *EMBO J.* **22**, 668–675 (2003).
8. Nangle, L. A., Motta, C. M. & Schimmel, P. Global effects of mistranslation from an editing defect in mammalian cells. *Chem. Biol.* **13**, 1091–1100 (2006).
9. Lee, J. W. et al. Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. *Nature* **443**, 50–55 (2006).
10. Lu, J., Bergert, M., Walther, A. & Suter, B. Double-sieving-defective aminoacyl-tRNA synthetase causes protein mistranslation and affects cellular physiology and development. *Nat. Commun.* **5**, 5650 (2014).
11. Liu, Y. et al. Deficiencies in tRNA synthetase editing activity cause cardioproteinopathy. *Proc. Natl Acad. Sci. USA* **111**, 17570–17575 (2014).
12. Calendar, R. & Berg, P. d-Tyrosyl RNA: formation, hydrolysis and utilization for protein synthesis. *J. Mol. Biol.* **26**, 39–54 (1967).
13. Pawar, K. I. et al. Role of d-aminoacyl-tRNA deacylase beyond chiral proofreading as a cellular defense against glycine mischarging by AlaRS. *eLife* **6**, e24001 (2017).
14. Calendar, R. d-tyrosyl-tRNA deacylase: a new function. *Trends Biochem. Sci.* **42**, 684–686 (2017).
15. Niwa, H., Yamamura, K. & Miyazaki, J. Efficient selection for high-expression transfectants with a novel eukaryotic vector. *Gene* **108**, 193–199 (1991).
16. Okabe, M., Ikawa, M., Kominami, K., Nakanishi, T. & Nishimune, Y. 'Green mice' as a source of ubiquitous green cells. *FEBS Lett.* **407**, 313–319 (1997).
17. Tsui, W. C. & Fersht, A. R. Probing the principles of amino acid selection using the alanyl-tRNA synthetase from *Escherichia coli*. *Nucleic Acids Res.* **9**, 4627–4637 (1981).
18. Yanagisawa, T., Sumida, T., Ishii, R., Takemoto, C. & Yokoyama, S. A paralog of lysyl-tRNA synthetase aminoacylates a conserved lysine residue in translation elongation factor P. *Nat. Struct. Mol. Biol.* **17**, 1136–1143 (2010).
19. He, X. D. et al. Sensing and transmitting intracellular amino acid signals through reversible lysine aminoacylations. *Cell Metab.* **27**, 151–166.e6 (2018).
20. Jakubowski, H. Protein homocysteinylation: possible mechanism underlying pathological consequences of elevated homocysteine levels. *FASEB J.* **13**, 2277–2283 (1999).
21. Jakubowski, H. Quality control in tRNA charging—editing of homocysteine. *Acta Biochim. Pol.* **58**, 149–163 (2011).
23. Beebe, K., Mock, M., Merriman, E. & Schimmel, P. Distinct domains of tRNA synthetase recognize the same base pair. *Nature* **451**, 90–93 (2008).
24. Chong, Y. E., Yang, X. L. & Schimmel, P. Natural homolog of tRNA synthetase editing domain rescues conditional lethality caused by mistranslation. *J. Biol. Chem.* **283**, 30073–30078 (2008).
25. Fersht, A. R. Editing mechanisms in protein synthesis. Rejection of valine by the isoleucyl-tRNA synthetase. *Biochemistry* **16**, 1025–1030 (1977).
26. Jakubowski, H. & Goldman, E. Editing of errors in selection of amino acids for protein synthesis. *Microbiol. Rev.* **56**, 412–429 (1992).
27. Ibba, M. & Söll, D. Quality control mechanisms during translation. *Science* **286**, 1893–1897 (1999).
28. van Ham, T. J., Breitling, R., Swertz, M. A. & Nollen, E. A. A. Neurodegenerative diseases: Lessons from genome-wide screens in small model organisms. *EMBO Mol. Med.* **1**, 360–370 (2009).
29. Chen, X. & Burgoyne, R. D. Identification of common genetic modifiers of neurodegenerative diseases from an integrative analysis of diverse genetic screens in model organisms. *BMC Genomics* **13**, 71 (2012).
30. Lenz, S., Karsten, P., Schulz, J. B. & Voigt, A. *Drosophila* as a screening tool to study human neurodegenerative diseases. *J. Neurochem.* **127**, 453–460 (2013).

Acknowledgements We thank K. Brown, J. Cook, T. Jucius and A. Kano for technical assistance, and the microinjection core at The Jackson Laboratory for transgenic mouse production. This work was supported by National Institutes of Health grants R01NS42613 to S.L.A., R01GM072528 to K.F., R01CA92577 to P.S., and a Fellowship from the National Foundation for Cancer Research to P.S. S.L.A. is an investigator of the Howard Hughes Medical Institute.

Reviewer information *Nature* thanks C. Francklyn and M. Justice for their contribution to the peer review of this work.

Author contributions S.L.A., K.F., P.S., M.T. and M.-N.V. designed experiments and wrote the paper. J.-W.L. performed gene mapping and in vivo transgenic rescue experiments, M.T. performed immunofluorescence and immunohistochemical experiments, tissue- and cell-based immunoprecipitations, cell fractionation, western blotting and RT–PCR studies, M.T. and J.-W.L. performed MEF studies, H.F. characterized the ANKRD16 antibody, M.-N.V. performed enzyme assays and bacterial rescue experiments, L.S. performed the protein modelling and deacylation experiments, L.S. and M.-N.V. performed direct immunoprecipitation studies, T.G.W. performed switchSENSE measurements, B.R. and Q.L. performed dipeptide assays under the guidance of K.F., and J.J.M. performed the mass spectrometry under the guidance of J.R.Y.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0137-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0137-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to P.S. or S.L.A.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Ethical compliance. All mouse studies were performed under the guidance of the University of California, San Diego and The Jackson Laboratory Animal Care and Use Committee in accordance with institutional and regulatory guidelines.

Mice. Genetic mapping of the *Msti* (modifier of sticky) was performed on N_1 and F_2 mice using polymorphic microsatellite and SNP markers (primer sequences are listed below). In brief, C57BL/6J (B6).*Aars*$^{sti/+}$ mice were intercrossed with CAST/E_i (CAST) mice and the resulting F_1 *Aars*$^{sti/+}$ mice were either backcrossed to B6.*Aars*$^{sti/+}$ to generate N_1 *Aars*$^{sti/sti}$ mice ($n = 1{,}249$) or intercrossed to generate F_2 *Aars*$^{sti/sti}$ mice ($n = 766$) for phenotypic assessment. Ataxia and Purkinje cell degeneration of N_1 or F_2 *Aars*$^{sti/sti}$ mice were assessed at three to four months of age. To generate *Ankrd16*$^{-/-}$*Aars*$^{sti/sti}$ embryos, timed matings between *Ankrd16*$^{-/-}$*Aars*$^{sti/+}$ and *Ankrd16*$^{-/-}$*Aars*$^{sti/+}$ mice were performed. The day that a vaginal plug was detected was defined as embryonic day 0.5 (E0.5).

To generate heterozygous congenic B6.CAST-*Msti* mice (*Msti*$^{CAST/B6}$), B6 and CAST mice were intercrossed and mice carrying the CAST allele at the *Msti* flanking markers (*D2Mit4* and *D2Mit427*) were backcrossed to B6 mice for seven generations. All transgenic mice were generated by pronuclear injection of B6 zygotes. Tg25L9 transgenic mice were generated by injection of a CAST-derived BAC clone (CH26-25L9). To generate transgenic TgPcp2 *Ankrd16* mice, the coding sequence of *Ankrd16*^{CAST} cDNA was inserted into the BamHI site of the fourth exon of the L7/*Pcp2* minigene[31]. To generate Tg-CAG *Ankrd16*-Myc transgenic mice, the stop codon of the *Ankrd16*^{CAST} cDNA was replaced with three copies of a *myc* epitope tag and cloned into the EcoRI site of the pCAGGS vector[15].

Generation of the targeted *Ankrd16* allele was performed by homologous recombination of the *Ankrd16* locus in R1 embryonic stem cells. Targeted embryonic stem cells were confirmed by Southern blotting. Cells were injected into B6J blastocysts to generate *Ankrd16*$^{neo/+}$ mice, which were backcrossed ten generations to B6J mice. Ubiquitous deletion of exon 2 of *Ankrd16* to generate the *Ankrd16*$^{+/-}$ allele was accomplished by crossing *Ankrd16*$^{neo/+}$ mice to B6.FVB-Tg(EIIa-cre) C5379Lmgd/J (The Jackson Laboratory, stock #003724). To generate a conditional *Ankrd16*$^{fl/+}$ allele, the neo cassette was removed by crossing *Ankrd16*$^{neo/+}$ mice to B6.129S4-*Gt(ROSA)26Sor*$^{tm1(FLP1)Dym}$/RainJ (The Jackson Laboratory, stock #009086). For conditional loss-of-function experiments, the following Cre-lines were used: B6.129-Tg(Pcp2-cre)2Mpin/J (The Jackson Laboratory, stock #004146), *En1*$^{tm2(cre)Wrst}$/J (The Jackson Laboratory, stock #007916), B6.Cg-Tg(Camk2a-cre) T29-1Stl/J (The Jackson Laboratory, stock #005359).

SNP markers are as follows: *D2SlacCA3* forward: 5′GCTAGAAAGATGCTGG TAATGGAA3′, *D2SlacCA3* reverse: 5′GGCTGGCTGTGTAAGCACAT3′; *D2SlacCA6* forward: 5′AACACACCATACCACACACACA3′, *D2SlacCA6* reverse: 5′TGGAGGTTTGCAAAGGAATC3′; *D2SlacCA8* forward: 5′ATAC CCACCCACATGTGACG3′, *D2SlacCA8* reverse: 5′GTACAAGATACTG AGAGTTGGT3′; *D2SlacCA10* forward: 5′AGGCTGGAGGACTGTGGGTT3′, *D2SlacCA10* reverse: 5′GCGATTGTCCATGGTGGCTT3′; *D2SlacCA11* forward: 5′CATCTGGTTTATGTGAATGCC3′, *D2SlacCA11* reverse: 5′CAGAGTGCTTTA CCTTTCTC3′; *D2SlacCA12* forward: 5′AATGGTCAGCCCTGGAAAC3′, *D2SlacCA12* reverse: 5′ACATTCAGCCTGCTTGGTCT3′; *D2SlacTAGA1* forward: 5′TGGATGAAGATTTTACAGAT3′, *D2SlacTAGA1* reverse: 5′ATCTATC TGTCTGTCTATCC3′; *Intr. 5 Ank16* forward: 5′GTAGGGT TGGAGAGATAGCC3′, *Intr. 5 Ank16* reverse: 5′TGTGAACACAGTAATAC AGGTGC3′.

Rotorod assay. Mice were placed on a rotarod linearly accelerating from 4 to 40 r.p.m. (type 7650, Ugo Basile). The latency to fall was averaged from four trials and repeated over the course of four consecutive test days. A test was considered over when mice fell, 360 s expired, or the mouse used gripping behaviour to remain on the rod. Rotarod experiments were performed with mice of both sexes and at least 15 mice per genotype (*Msti*$^{B6/B6}$*Aars*$^{+/+}$, $n = 15$; *Msti*$^{B6/B6}$*Aars*$^{sti/sti}$, $n = 16$; *Msti*$^{CAST/B6}$*Aars*$^{sti/sti}$, $n = 15$). Data represent mean ± s.d. and were analysed by two-way ANOVA using Tukey's multiple comparison tests.

Histology and immunohistochemistry. Anesthetized mice were transcardially perfused with 4% paraformaldehyde (PFA, for immunofluorescence and histology) or acetic acid/methanol (for calbindin-D28 immunohistochemistry and immunofluorescence; for example, aggregate staining). Tissues were post-fixed overnight and embedded in paraffin. For histological analysis, sections were deparaffinized, rehydrated and stained with cresyl violet according to standard procedures. For immunofluorescence, antigen retrieval on paraformaldehyde-fixed sections was performed by microwaving sections in 0.01 M sodium citrate buffer (pH 6.0, 0.05% Tween-20) three times for 3 min each. Paraformaldehyde—or acetic acid/methanol-fixed sections were incubated with the following primary antibodies: rabbit anti-ubiquitin (DAKO, # Z0458, 1:1,000), mouse anti-ubiquitin (Cell Signaling, #3936, 1:350), guinea pig anti-p62 (ARP, 03-GP62-C, 1:300), rabbit anti-Calbindin D28 (Swant, CB38, 1:300), and guinea pig anti-ANKRD16 (1:200). Guinea pig or rabbit anti-ANKRD16 antibodies were custom-made against recombinant mouse ANKRD16 (1–154 aa) and affinity purified. Detection of primary antibodies was performed with Alexa Fluor-488, -555 or -350 conjugated goat anti-mouse, -rabbit or -guinea pig secondary antibodies (Invitrogen). Sections were counterstained with DAPI, and autofluorescence was quenched with Sudan Black.

For aggregate quantification, the total number of Purkinje cells or hippocampal pyramidal neurons and ubiquitin- or p62-positive neurons were quantified using three sections spaced 100 μm apart per mouse. The percentage of cells with aggregates ± s.d. at different time points were compared between genotypes using multiple *t*-tests and corrected for multiple comparisons using the Holm–Sidak correction (Fig. 1c and Extended Data Fig. 9a). All histological analyses were performed with at least three mice of each genotype and time point, using mice of both sexes.

Constructs. For the generation of the various *Aars* constructs, a full-length mouse *Aars* clone (Open Biosystems/GE Dharmacon) was used as a template. The wild-type mouse *Aars* sequence was subcloned and inserted into a mammalian expression vector (pCMV3, Agilent Technologies) via restriction digestion. To generate the mutant AlaRS(A734E) construct, the mutant editing domain of *Aars* was amplified by PCR from total brain cDNA of B6.*Aars*$^{sti/sti}$, TOPO subcloned (TOPO TA cloning kit, ThermoFisher, K450002), and the wild-type editing domain of pCMV3-Aars was replaced with the mutant editing domain (pCMV3-AlaRS(A734E)) via restriction digestion. In addition, wild-type and mutant *Aars* (pCMV3) were digested and inserted into pCruz HA (C) vector (HA, human influenza haemagglutinin; Santa Cruz, sc-5045). To generate AlaRS deletion constructs (pCMV3-AAD-Aars, pCMV3-ΔAAD-Aars, pCMV3-ΔAAD-AlaRS(A734E)), individual domains of AlaRS were amplified by PCR using either full-length wild-type or mutant *Aars* constructs, TOPO subcloned, and inserted into pCMV3 vector via restriction digestion. The TgCAG *Ankrd16*-Myc construct was used to amplify *Ankrd16* by PCR, whereas *Ankrd29* was amplified from total brain B6J cDNA. PCR products were TOPO subcloned and then inserted into pCMV3 via restriction digestion. QuickChange II Site-Directed Mutagenesis kit (Agilent, #20052) was used to mutate lysines 102, 135 and 165 in ANKRD16 to arginines (ANKRD16(3×R)).

Cell culture. Mouse embryonic fibroblasts (MEFs) were prepared by standard procedures[32]. HEK293T (American Type Cell Culture, not authenticated nor tested for mycoplasma contamination) and MEF cells were maintained in Dulbecco's modified Eagle's medium (Sigma) with GlutaMAX (Gibco) and 10% fetal bovine serum at 37 °C in 5% CO_2. HEK293T cells were transfected with various constructs as listed. Transfections were performed using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol and cultured for 48 h before co-IP and western blot analysis.

For the serine toxicity assay, freshly isolated MEFs were cultured in increasing concentrations of serine for 24 h; MEFs were collected and stained with propidium iodide. The percentage of cell death was measured by fluorescence-activated cell sorting ($n = 3$, Fig. 3d). Data represent mean ± s.d. and were analysed by two-way ANOVA using Tukey's multiple comparison tests. For fluorescence microscopy ($n = 4$, Fig. 4g), *Aars*$^{sti/sti}$ embryonic fibroblasts were co-transfected with phr-*GFP* II and either ANKRD16 Flag or ANKRD16(3×R)–Flag using the P4 Primary Cell 4D-Nucleofector X Kit (Lonza), with around 50% transfection and around 95% co-transfection efficiency. 12 h after transfection and seeding, cells were washed twice with PBS and media with either 0.4 mM or 40 mM serine was added for 24 h. Subsequently, cells were stained with propidium iodide (Invitrogen, P21493, 250 ng ml^{-1}) and Hoechst 33342 (Life Technologies, H3570, 5 μg ml^{-1}) to determine cell death. The number of cells that were propidium iodide- and GFP-positive out of the total GFP-positive cells was determined. About 1,000 cells were counted per genotype and condition. Cell death was expressed as the percentage increase in cell death between low and high serine conditions. Data represent mean ± s.e.m. and were analysed by one-way ANOVA using Tukey's multiple comparison tests.

Cell fractionation. Brains from six-week-old mice were isolated and immediately processed for subcellular fractionation as previously described[33,34]. In brief, brains were excised, immersed in ice-cold PBS, and washed three times. After washing, brain tissue was minced into smaller pieces and homogenized in buffer (0.25 M sucrose, 10 mM HEPES, pH 7.5, protease inhibitor cocktail (Roche)) using a Potter-Elvehjem Tissue Grinder. The homogenate was centrifuged at 1,000*g* for 10 min to generate a crude cytoplasmic fraction and nuclear fraction for further purification by ultracentrifugation. The supernatant (crude cytoplasmic fraction) from four brains was pooled to purify the mitochondrial, rough endoplasmic reticulum and cytoplasmic fractions, and the pellet (crude nuclear fraction) from two brains was used to purify the nuclear fraction. Eventually, purified pellets (mitochondria, nucleus, rough endoplasmic reticulum) were lysed (4 °C, 30 min) in lysis buffer (50 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS) and proteins from the cytoplasmic fraction were concentrated by ultrafiltration (Millipore, Amicon Ultra-15, #UFC900308). Equal amounts of total protein from each cellular fraction were subject to electrophoresis on SDS–PAGE gels.

Reverse transcription, quantitative PCR, and genomic PCR analysis. Cerebella from various inbred strains were isolated and immediately frozen in liquid

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

nitrogen. Total RNA was extracted with TRIzol reagent (Life Technologies). cDNA synthesis was performed on DNase-treated (DNA-free DNA Removal Kit, Life Technologies AM1906) total RNA using oligo(dT) primers and the SuperScript III First-Strand Synthesis System (Invitrogen/Life Technologies). Quantitative RT–PCRs (qRT) were performed using iQ SYBR Green Supermix (Bio-Rad) and an iQ5 Multicolor Real-Time PCR Detection System (Bio-Rad). Reactions were performed with primers either complementary to the CAST or to the B6 *Ankrd16* sequence. Both primer sets showed similar results; results using B6 primers are shown in Fig. 2d. Expression of *Ankrd16* transcript was normalized to *18S* rRNA using the $2^{-\Delta\Delta CT}$ method[35], expressed as fold change ± standard error of the mean (s.e.m.) relative to control (C57BL/6J) and analysed by multiple *t*-tests using Holm Sidak comparison tests ($n = 3$). *Ankrd16* quantitative and conventional RT–PCR analysis was performed on 50 ng of cerebellar cDNA generated from one- or two-month-old mice, respectively.

To confirm non-synonymous SNPs of candidate genes, PCRs were performed on genomic DNA from various inbred strains. PCR products were purified (Agencourt AMPure XP—PCR Purification, Beckman Coulter) and subjected to sequencing.

The RT–PCR primers used are as follows: 5′UTR *Ankrd16* forward: 5′GTCTT CCTCCTACTTTTGTCCA3′, Exon2 *Ankrd16* reverse: 5′CAAG G TTCTTCCTT GTGCAC3′; Exon1 *Ankrd16* forward: 5′TCGTGGACTC CTTGAAGAAG3′, Exon3 *Ankrd16* reverse: 5′TTCCAAACAGCCGTGCATTG3′, Exon2 *Ankrd16* forward: 5′TGATCCTCCGGTACTTGCT3′, Exon5 *Ankrd16* reverse: 5′ATCTAC ATCGATGCCAAGACC3′, Exon4 *Ankrd16* forward: 5′GCTGCTCCTT GAACAGCATAA3′, Exon6 *Ankrd16* reverse: 5′CACCCAAGGACAACAG AGTT3′, Exon5 *Ankrd16* forward: 5′CAGCACTTCACTATGCAGCA3′, Exon7 *Ankrd16* reverse: 5′CATG GTCAAAGTCCTGAAGGA3′, Exon6 *Ankrd16* forward: 5′GTGCCGACATCAACTCTACA3′, 3′UTR *Ankrd16* reverse: 5′TGAACGGCCTT GAACTCCAT3′; Exon3 *Ankrd16* forward: 5′GCTGTTTGGAAGCAGTCCA3′, Exon5′ *Ankrd16* reverse: 5′CCAAGGAAGAGCTTACGCTAT3′.

qRT–PCR primers: Exon 5/6 (B6) forward: 5′TGCAGC*A*AAGGAAGGAC AGA3′, or Exon 5/6 *(*CAST) forward: 5′TGCAGCGAAGGAAGGACAGA3′, Exon7 *Ankrd16* reverse: 5′GTAGGAGGAGCCTGGTGCAA3′; 18S RNA forward, 5′ GAGGGAGCCTGAGAAACG G 3′, 18S RNA reverse: 5′GTCGGGAG TGGGTAATTTGC3′.

Non-synonymous SNP primers (*Msti* candidate genes): *Il2ra*(T101S/Q106H) forward: 5′TCCCATATTCTCCAGCCCAT3′, *Il2ra*(T101S/Q106H) reverse: 5′CACCTGCCTGTTGAGAACA3′; *Il2ra*(S143R) forward: 5′TGCTGT GTCTTCCCAAACTT3′, *Il12ra*(S143R) reverse: 5′CCCGTTTTC CCACACTTCAT3′; *Il2ra*(G158S) forward: 5′CCACCTCCTTGGAAACATGA3′, *Il2ra*(G158S) reverse: 5′TCGGTGGTGTTCTCTTTCATC3′; *Il2ra*(M225T) forward: 5′GCAGTCCTATGCTAGCCATAA3′, *Il2ra*(M225T) reverse: 5′CGTCTTTGCATGCTTCACC3′; *Il2ra*(K236Q/M233V) forward: 5′TTAG TGGCCGTTTCAACCAG3′, *Il2ra*(K236Q/M233V) reverse: 5′CTTGTGCAG TCCTATGCTAGC3′; *Fbxo18*(H15Q) forward: 5′CCCTATTCCGTCCT TTTGTTC3′, *Fbxo18*(H15Q) reverse: 5′ATGAGACGGTTTAAGCGGA A3′; *Fbox18*(P369/376A) forward: 5′ATTAACGTCTGGGCCTTGGT3′, *Fbxo18*(P369/376A) reverse: 5′TGCTGATGTTAATCCCCTTCTC3′. *Ank16* (E335A/T346M) forward: 5′ATCTGGCCTGCGCAGGTCA3′, *Ank16*(E335A/ T346M) reverse: 5′CTTGAACTCCATCGCCTC3′.

Co-immunoprecipitation experiments. Brain or liver tissue from four-week-old Tg-CAG *Ankrd16*-Myc or non-transgenic mice were isolated and immediately used or frozen in liquid nitrogen. Whole protein lysate from either tissue samples or HEK293T cells was extracted with solubilization buffer (20 mM Tris-HCl (pH 7.4), 137 mM NaCl, 0.1% Triton X-100, Protease Inhibitor Cocktail (Roche)) using a Potter-Elvehjem Tissue Grinder (for tissue samples). Protein extraction was performed at 4 °C for either 45 min (tissue samples) or 30 min (cell culture samples) under gentle agitation. The lysate was centrifuged at 16,000*g* at 4 °C for either 30 min (tissue samples) or 20 min (cell culture samples). The lysate was pre-cleared in two steps at 4 °C under gentle agitation, first with Dynabeads Protein G (Dynabeads Protein G Immunoprecipitation Kit, Invitrogen, 10007D) for 90 min followed by clearing with normal IgG complex for 90 min (corresponding to the species of IP antibody, Santa Cruz, mouse IgG (sc-2025), or goat IgG (sc-2028)) that was pre-immobilized on Dynabeads Protein G. Subsequently, pre-cleared lysates were incubated with pre-immobilized Dynabeads Protein G with antibody (mouse anti-Myc, Abcam, ab18185; mouse anti-Flag, Sigma, clone M2, F1804; or goat anti-HA, Santa Cruz, HA probe Y-11, sc-805-G) for 3 to 5 h at 4 °C under gentle agitation. After washing with wash buffer (20 mM Tris-HCl (pH 7.4), 137 mM NaCl, 0.5% Triton X-100 (tissue samples) or 20 mM Tris-HCl (pH 7.4), 300 mM NaCl, 0.1% Triton X-100 (cell culture samples)), proteins were eluted by incubation with SDS–PAGE sample buffer for western blotting.

Mass spectrometry. Tissue lysates were prepared as described for co-immunoprecipitation and proteins were eluted twice by incubation with 2% SDS for 5 min at 90 °C. After elution, proteins were recovered in 100 mM Tris buffer (pH 7.4), and purified using chloroform/methanol precipitation. Protein pellets were dried using a speed vac and dissolved in 30 µl of 20 mM Tris buffer (137 mM NaCl, pH 7.4). The two elutions for each sample were combined (total volume 60 µl), then mixed with 33.5 µl of 50 mM NH_4HCO_3 and 1 µl of 0.5 M tris(2-carboxyethyl)phosphine hydrochloride (Thermo Scientific, #20490). Samples were incubated for 20 min, occasionally vortexed, mixed with 2.7 µl of 0.55 M iodoacetamide, incubated for 15 min in the dark, and mixed with 1 µl of 1% ProteaseMax (Promega, V2071) to enhance the subsequent digestion. Trypsin digests were performed at 37 °C overnight following the instructions for 'solution digestion' using Trypsin Singles (2 µg of trypsin per sample, Proteomics Grade, Sigma, T7575), mixed with 5% formic acid (final concentration 0.5%), centrifuged for 30 min at 13,000*g*, and the supernatant was collected for analysis by mass spectrometry.

For mass spectrometric analysis, a four-step MudPIT analysis was performed using an Accela pump and a Thermo LTQ-XL (ThermoFisher Scientific) using an electrospray stage built in-house[36]. Tandem mass spectra were extracted from raw files using RawExtract 1.9.9[37] and were searched against a Uniprot *Mus musculus* database with reversed sequences using ProLuCID[38]. The search space included all fully-tryptic and half-tryptic peptide candidates with a fixed modification of 57.02146 C. Proteins were considered putative interaction candidates if they were detected with at least five peptides (total number of peptides). Proteins detected in both transgenic (experimental sample) and non-transgenic (negative sample) tissue were considered non-specific and excluded as putative interaction candidates.

Western blotting. Proteins were extracted with RIPA buffer (including Proteinase Inhibitor cocktail, Roche) or solubilization buffer (see above) when used for co-IP experiments, and were resolved on SDS–PAGE gels before transfer to PVDF membranes (GE Healthcare Life Sciences, #10600023) using a tank blotting apparatus (BioRad). After blocking in 5% non-fat dry milk (Cell Signaling, #9999S), blots were probed with primary antibodies at 4 °C overnight: mouse anti-Myc (Abcam, ab18185, 1:1,000), goat anti-Myc (Abcam, ab9132, 1:1,000), rabbit anti-Myc (Abcam, ab9106, 1:2,000), guinea pig anti-ANKRD16 (1:150), rabbit anti-ANKRD16 (1:400), mouse anti-AlaRS (A-6, Santa Cruz, sc-165990, 1:100), rabbit anti-AlaRS (Sigma, HPA040870, 1:500), rabbit anti-GAPDH (Cell Signaling, #2118, 1:6,500), mouse anti-Flag (Sigma, M2 clone, #F1804, 1:5,000), rabbit anti-Flag (Sigma, F-7425, 1:5,000), mouse anti-HA (Covance, MMS-101P, 1:2,000), rabbit anti-Histone 3-HRP (2 h incubation at room temperature, Cell Signaling, #5192, 1:500), rabbit anti-COX IV (Cell Signaling, #4850, 1:5,000), rabbit anti-Grp78 (StressGen, SPA-826, 1:500), goat anti-Sec61β (Santa Cruz, sc-27694, 1:500) and followed by incubation with HRP-conjugated secondary antibodies for 2 h at room temperature: goat anti-rabbit IgG (BioRad, #170-6515), goat anti-mouse IgG (BioRad, #170-6516), bovine anti-goat IgG (Santa Cruz, sc-2384), donkey anti-goat IgG (Santa Cruz, sc-2056), goat anti-guinea pig IgG (Santa Cruz, sc-2903) or donkey anti-guinea pig IgG (Millipore, AP193P). Signals were detected with SuperSignal West Pico Chemiluminescent substrate (Thermo Scientific, #34080). Quantification of protein levels (cerebellum, C57BL/6J, CAST/E_i, $Aars^{sti/sti}$, $Aars^{sti/sti}Msti^{CAST/CAST}$ and $Msti^{CAST/B6}$, three- to four-week-old mice, $n = 3$) and tissue expression (C57BL/6J, $Msti^{CAST/B6}$ and $Msti^{CAST/CAST}$, six-week-old mice) of ANKRD16 was determined, normalized to *GAPDH*, and expressed as a fold change ± s.d. relative to $Aars^{sti/sti}$ and analysed by ordinary one-way ANOVA (Tukey-test). Quantification of binding affinity was determined by normalizing the immunoprecipitation signal to the input signal of AlaRS.

Protein expression and purification. Histidine-tagged human or mouse AlaRS (full length and truncations (AlaRS-ΔAAD has an N-terminal SUMO tag)), *E. coli* AlaRS, human TyrRS, human TrpRS, mouse ANKRD16 and mouse ANKRD29 were expressed in *E. coli* and purified by Ni-NTA resin (Qiagen), followed by HiLoad 16/60 Superdex 200 prep grade column (GE Healthcare). Glutathione *S*-transferase (GST)-tagged mouse ANKRD16 was expressed in *E. coli* and purified by glutathione-Sepharose 4B beads (GE Healthcare), followed by a HiLoad 16/60 Superdex 200 prep grade column. The final purified proteins were stored in the 20 mM HEPES (pH 7.5), 200 mM NaCl, and 1 mM dithiothreitol (DTT).

In vitro transcription of $tRNA^{Ala}$. The DNA template containing a T7 promoter and a $tRNA^{Ala}$ gene was synthesized by PCR of overlapping oligonucleotides. The transcription reaction was performed in 40 mM Tris-HCl (pH 8.0), 25 mM NaCl, 20 mM $MgCl_2$, 2 µg ml^{-1} pyrophosphatase, 0.1 mg ml^{-1} bovine serum albumin (BSA), 5 mM DTT, 20 mM NTPs with T7 polymerase and the DNA template at 37 °C for 2 h. The tRNA transcript was purified by phenol–chloroform extraction. The purified tRNA was annealed by heating to 95 °C for 3 min and then slowly cooled to room temperature, with addition of 1 mM $MgCl_2$ at 55 °C.

Dipeptide translation experiments. Ribosomal subunits (from *E. coli* MRE600) and EF-Tu were purified as described[39,40]. The model mRNA used is identical to m291[41], except with GCU as codon 2, and was made by in vitro transcription followed by PAGE purification. Native *E. coli* $tRNA^{fMet}$ was purchased (Chemical Block Ltd) and charged and formylated as described[42]. A transcript corresponding to human $tRNA^{Ala}$ (AGC) was generated in vitro, purified by PAGE and charged with alanine and serine[42] using AlaRS and AlaRS(1–455), respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

70S initiation complex was formed by incubating heat-activated 30S subunits (0.2 μM), mRNA (1 μM), formyl-[^{35}S]-Met-tRNA^{fMet} (0.2 μM), 50S subunits (0.6 μM) and 1 mM GTP in buffer A (50 mM K-HEPES pH 7.6, 10 mM KCl, 100 mM NH_4Cl, 10 mM $MgCl_2$, 1 mM DTT) at 37 °C for 30 min. Charged tRNA was used to form the ternary complex by incubating either Ser-tRNA^{Ala} or Ala-tRNA^{Ala} (1 μM) in buffer A with EF-Tu (4 μM), GTP (1 mM), phosphoenol pyruvate (2 mM), and pyruvate kinase (Sigma; 50 μg ml^{-1}) at 37 °C for 15 min. The ternary complex was used in reactions at 1 μM. ANKRD16 was purified as above and used in reactions at 4 μM.

For experiments that begin with deacylated tRNA, tRNA^{Ala} (5 μM) was pre-equilibrated at 37 °C with EF-Tu (8 μM), amino acid (for example, 100 μM alanine (Fig. 4a), 500 μM serine and 1 μM alanine (Fig. 4b) or 500 μM serine (Extended Data Fig. 3c)), GTP (1 mM), ATP (10 mM), phosphoenol pyruvate (2 mM), pyruvate kinase (50 μg ml^{-1}), pyrophosphatase (Roche; 8 ng μl^{-1}), and BSA (4 μM) in buffer A, and AlaRS (4 μM) was added to initiate the reaction. Peptide products formed upon reaction of 70S initiation complex and the ternary complex were resolved by electrophoretic thin layer chromatography (eTLC)[43] and quantified using a Typhoon FLA9000 phosphoimager (GE Healthcare). Apparent rates were obtained by fitting the kinetic data to a single exponential equation.

***E. coli* strain construction and halo assay.** Editing-defective AlaRS(C666A/Q584H) *E. coli* was prepared and the halo assay performed as described[24]. In brief, *E. coli* were transformed with empty pBAD33/21 vector or pBAD33/21 containing mouse wild-type ANKRD16, ANKRD16(3×R) or mouse ANKRD29. The resulting strains were grown overnight in M9 minimal media with 0.4% glycerol, 0.002% L-arabinose, 0.01 mg ml^{-1} thiamine and the following antibiotics: kanamycin (50 μg ml^{-1}), ampicillin (100 μg ml^{-1}), and chloramphenicol (30 μg ml^{-1}). Five microlitres of an overnight culture were diluted (1:100) and spread on M9-Kan/Amp/Cm plates. One hundred microlitres of 1 M L-serine were added to a well cut into the centre of the plate and allowed to diffuse into the agar, creating a radial gradient. Plates were incubated at 37 °C and imaged after 48 h.

SwitchSENSE measurements. Wild-type AlaRS, AlaRS(A734E) and AlaRS(1–455) were coupled to the 5′ end of a 48-mer ssDNA complementary nanolever sequence (cNL-B48) using the Amine Coupling Kit 1 (Dynamic Biosensors), and the products were purified using an ÄKTA start system (GE Life Sciences) equipped with an Anion Exchange Column (Dynamic Biosensors). The cNL-B48 conjugates were then hybridized to a 3′ red-fluorescent labelled 48mer ssDNA nanolever tethered to a biochip electrode. The analyte, wild-type or mutant ANKRD16, was injected onto the electrode in three different concentrations (0.63, 1.25, and 2.50 or 5.00 μM) at a constant flow rate of 50 μl min^{-1} for 3.5 min and dissociation was observed at a buffer flow rate of 200 μl min^{-1} for 6 min. Measurements were performed on a DRX2 instrument (Dynamic Biosensors) and switchSENSE technology was employed as previously described[44]. Molecular interactions were recorded at constant negative potential (−0.1 V) on the basis of the changes in the red fluorescence signal upon analyte binding. Kinetic curves were referenced to the corresponding buffer control and kinetic values were determined by global monophasic fitting using switchANALYSIS software (Dynamic Biosensors).

Aminoacylation and misacylation assays. Aminoacylation assays were performed as described[7]. In brief, assays were carried out at room temperature using either mouse AlaRS(A734E) alone (50 nM, Extended Data Fig. 3g or 2 μM, Extended Data Fig. 4e) or a mixture of AlaRS(A734E) and 10 μM ANKRD16 with 50 mM HEPES (pH 7.5), 20 mM KCl, 4 mM ATP, 5 mM $MgCl_2$, 2 mM DTT, 4 μg ml^{-1} of inorganic pyrophosphatase (Roche), 20 μM [^{3}H]-L-alanine, 2 μM BSA and 10 μM in vitro transcript corresponding to human tRNA^{Ala}.

Misacylation assays were performed at 25 °C in reactions with 5 μM AlaRS(A734E) alone or a mixture of 5 μM AlaRS(A734E) and 10 μM ANKRD16 with 50 mM HEPES (pH 7.5), 50 mM KCl, 4 mM ATP, 10 mM $MgCl_2$, 5 mM DTT, 4 μg ml^{-1} of inorganic pyrophosphatase (Roche), 50 μM [^{3}H]-L-serine, 2 μM BSA and 10 μM in vitro transcript corresponding to human tRNA^{Ala}.

Serine- or alanine-linked ANKRD16 (aa-linked ANKRD16) and aminoacylated-tRNA^{Ala} (aa-linked tRNA^{Ala}) were determined based on aminoacylation reactions with [^{3}H]alanine or [^{3}H]serine using 5 μM AlaRS(A734E) in the presence of 10 μM tRNA^{Ala} and 10 μM ANKRD16. The reaction was incubated for 1 h, then half of the reaction was treated with alkaline pH (Na_2CO_3) for 20 min to hydrolyse acylated tRNA^{Ala} (leaving aa-linked ANKRD16), followed by trichloroacetic acid (TCA) precipitation. Aminoacylated tRNA^{Ala} was the difference between total apparent aminoacylated tRNA (no alkaline hydrolysis) and aa-linked ANKRD16 (after alkaline hydrolysis). Data represent mean ± s.d. and were analysed by a two-tailed Student's *t*-test.

ATP hydrolysis assays. ATPase assays were performed as previously described[45]. Assays measuring the hydrolysis of misactivated serine (Ser-AMP) were carried out at room temperature in a reaction with either 500 nM mouse AlaRS alone, or a mixture of 500 nM AlaRS and 10 μM wild-type ANKRD16, mutant ANKRD16, or ANKRD29 proteins. Reactions also contained 100 mM HEPES (pH 7.5), 40 mM KCl, 10 mM $MgCl_2$, 0.5 mM ATP, 0.5 μM [α-^{32}P]ATP, 10 units per ml inorganic pyrophosphatase, 2 mM DTT and 200 mM serine. Two microlitre aliquots of the editing reaction were taken at appropriate time points and quenched in 8 μl 200 mM sodium acetate (pH 5.0), and then 1 μl was spotted on a polyethyleneimine thin-layer chromatography cellulose plate (EMD Millipore) and run in a solution containing 0.1 M ammonium acetate and 5% glacial acetic acid to separate [α-^{32}P]ATP, [^{32}P]AMP and Ser[^{32}P]AMP. Results were imaged using a phosphoimager. The percentage of hydrolysed AMP (ATPase activity) was quantified with the Molecular Dynamics Image Quant software. Data represent mean ± s.d. and were analysed by one-way ANOVA using Tukey's multiple comparison tests.

Far-UV circular dichroism. Circular dichroism spectra were obtained with a CD Spectrometer 400 (Aviv Biomedical, Inc.). Protein samples were prepared at 10 μM in PBS and measurements were scanned from 190 to 260 nm in increments of 0.5 nm at 20 °C. Three independent scans were acquired and averaged for each protein sample.

Thermal shift assays. Thermal shift assays were performed using the Protein Thermal Shift Dye Kit (ThermoFisher Scientific) according to the manufacturer's instructions and run on a StepOnePlus 96 Real Time Cycler (Applied Biosystems). In brief, a reaction of 20 μl containing 5.0 μl thermal shift buffer, 12.5 μl of 10 μM wild-type ANKRD16 or ANKRD16(3×R) and 2.5 μl of 8X thermal shift dye were mixed in a 96-well Optical Reaction Plate (Applied Biosystems). The plate was sealed and heated in the StepOnePlus Real-Time PCR from 25 to 90 °C in increments of 0.15 °C. The measurements were performed in four replicates and analysed using the Protein Thermal Shift software (ThermoFisher Scientific).

Deacylation assay. The aminoacylation assays were performed at room temperature with 50 mM HEPES (pH 7.5), 20 mM KCl, 5 mM $MgCl_2$, 4 mM ATP, 2 mM DTT, 4 μg ml^{-1} pyrophosphatase, 20 μM cold L-serine, 2.68 μM [^{3}H]serine (1 mCi ml^{-1}) as the assay solution. 20 μM tRNA^{Ala} transcript was mixed with the assay solution and the reaction was initiated by adding 5 μM AlaRS protein into the mixture. After 30 min, a phenol–chloroform (pH 5.0) extraction was performed, and the product [^{3}H]serine -tRNA^{Ala} was used as the substrate of the deacylation assay. [^{3}H]serine -tRNA^{Ala} was incubated at room temperature with 100 nM AlaRS (wild-type AlaRS and AlaRS(A734E)) and ANKRD16 protein in assay buffer (50 mM HEPES (pH 7.5), 20 mM KCl, 10 mM $MgCl_2$, 5 mM DTT). At varying time intervals, 5 μl aliquots were applied to a MultiScreen 96-well filter plate (0.45 μm pore size hydrophobic, low protein-binding membrane, Merck Millipore Ltd), which was pre-wetted with a quenching solution containing 0.5 mg ml^{-1} DNA and 100 mM EDTA in 300 mM NaOAc (pH 3.0). After all time points were collected, 100 μl of 20% TCA was added to precipitate the nucleic acids. The plate was then washed four times with 200 μl of 5% TCA containing 100 mM cold serine, then once with 95% ethanol. On drying after completion of the washing steps, 70 μl of 100 mM NaOH was added to elute the tRNA, which was then centrifuged into a 96-well flexible PET microplate (PerkinElmer) with 150 μl of Supermix scintillation cocktail (PerkinElmer). After mixing, the radioactivity in each well of the plate was counted using the 1450 LSC & Luminescence Counter (PerkinElmer).

ATP-pyrophosphate exchange assays. The ATP-pyrophosphate exchange assays were performed as previously described[46]. In brief, a reaction mixture of either 250 nM mouse AlaRS(A734E) alone or a mixture of 250 nM AlaRS(A734E) and 10 μM ANKRD16 was incubated at room temperature with 100 mM Tris-HCl (pH 7.5), 10 mM KF, 10 mM $MgCl_2$, 1 mM ATP, 2 mM DTT, 0.1 mg ml^{-1} BSA, 20 mM L-serine, and 0.5 mM sodium [^{32}P]pyrophosphate.

Filter binding of EF-Tu with aminoacyl-tRNA^{Ala}. The nitrocellulose filter-binding assays were performed as described[47], with some modifications. Aminoacyl-tRNA^{Ala} was generated in 200 μl aminoacylation reactions containing either 500 nM AlaRS(A734E) alone or a mixture of 500 nM AlaRS(A734E) and 10 μM ANKRD16, in reaction buffer with 50 mM HEPES (pH 7.5), 50 mM KCl, 4 mM ATP, 10 mM $MgCl_2$, 5 mM DTT, 4 μg ml^{-1} of inorganic pyrophosphatase (Roche), 50 μM L-alanine or L-serine, 2 μM BSA, and 10 μM in vitro transcript corresponding to human tRNA^{Ala}. Reactions were incubated for 30 min at room temperature. EF-Tu was loaded with GTP in a 200 μl reaction of 2 μM EF-Tu, 3 μM [^{3}H]GTP, 50 mM Tris-HCl (pH 7.4), 50 mM NH_4Cl, 10 mM $MgCl_2$, 3.75 mM phosphoenolpyruvate, 0.05 μg μl^{-1} pyruvate kinase, which was incubated for 2 min at 37 °C and then placed on ice. An equal volume (200 μl) of the aminoacylation reaction mixture or non-acylated tRNA^{Ala} was added to the [^{3}H]GTP-EF-Tu reaction mixture, incubated for 2 min on ice, and terminated by the addition of 2 ml of a cold wash buffer containing 10 mM Tris-HCl (pH 7.4), 50 mM NH_4Cl and 10 mM $MgCl_2$. The reaction mixture was filtered through a nitrocellulose filter, and the filter was washed with 15 ml of cold wash buffer. The filter was dissolved in a scintillation fluid and counted for radioactivity. The percentage of EF-Tu retained in the membrane was normalized to that retained in the non-acylated tRNA^{Ala} reaction.

GST pull-down assay. GST pull-down experiments were performed in buffer containing 20 mM of HEPES (pH 7.5), 200 mM NaCl, 0.5 mM EDTA, 1 mM DTT, 10% glycerol, and 0.1% Triton X-100. Equal amounts of GST-ANKRD16 proteins were incubated with histidine-tagged mouse or human AlaRS (full length and

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

truncations), *E. coli* AlaRS, or human TyrRS or TrpRS proteins for 2 h in 4 °C and pulled down by Glutathione-Sepharose 4B beads (GE Healthcare). After SDS–PAGE, the proteins were transferred to PVDF membranes using the iBlot Dry Blotting System (Invitrogen). The membranes were blocked for 1 h with 5% non-fat dry milk in Tris-buffered saline with Tween-20. Anti-GST and anti-His-tag antibodies were purchased from Cell Signaling. After incubation with primary antibodies, the membranes were washed and incubated with HRP-conjugated anti-mouse or anti-rabbit secondary antibodies (Cell Signaling), followed by detection using enhanced chemiluminescence western-blotting substrate (Thermo Scientific).

Molecular modelling. The mouse ANKRD16 and human AlaRS complex model was obtained through the Patchdock server and was optimized by manual adjustments. The mouse ANKRD16 and human AlaRS models were generated by the Swiss model server. The human full-length AlaRS model was built on the basis of the templates of the human AlaRS structure (PDB 5KNN[48]), *Archaeoglobus fulgidus* AlaRS and $tRNA^{Ala}$ complex (PDB 3WQY[49]). AlaRS was assigned as the receptor and ANKRD16 was assigned as the ligand. Molecular visualization and analysis were performed with PyMOL.

Mass spectrometry for serinylation of ANKRD16. To identify amino acid residues that were serinylated, misacylation assays were performed at 25 °C for 1 h in a reaction mixture of 5 μM AlaRS(A734E) and 10 μM ANKRD16 in 50 mM HEPES (pH 7.5), 50 mM KCl, 10 mM $MgCl_2$, 5 mM tris(2-carboxyethyl) phosphine, 4 μg ml^{-1} inorganic pyrophosphatase (Roche), 10 mM L-serine, 2 μM BSA and 10 μM human $tRNA^{Ala}$ in the presence or in the absence of 4 mM ATP. Proteins were reduced with 5 mM tris-(2-carboxyethyl)-phosphine hydrochloride (Sigma, C4706) and alkylated with 55 mM 2-chloroacetamide (Fluka Analytical, #22790). Proteins were digested for 18 h at 37 °C in 2 M urea, 100 mM Tris (pH 8.5), and 1 mM $CaCl_2$ with 2 μg trypsin (Promega, V5111). Five-step MudPIT analysis was performed using an Easy-nLC (ThermoFisher Scientific) and a Q Exactive (ThermoFisher Scientific) using an electrospray stage built in-house[36]. Tandem mass spectra were extracted from raw files using RawConverter[50] and were searched against a Uniprot *Mus musculus* database with reversed sequences using ProLuCID[38,51]. The search space included all fully tryptic peptide candidates with a fixed modification of 57.02146 C and differential modification of serine (87.03203) or alanine (71.03711) on lysine.

Statistics and reproducibility. No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to genotype during experiments. *P* values were computed in GraphPad Prism using Student's *t*-test (two-tailed), one-way or two-way ANOVA and corrected for multiple comparisons when applicable as indicated in the figure legends.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Source Data for quantifications mentioned either in the text or shown in graphs plotted in Figs. 1–4 and Extended Data Figs. 1, 3, 4, 5, 7, and 9 are available in the online version of this paper.

22. Kimmel, R. A. et al. Two lineage boundaries coordinate vertebrate apical ectodermal ridge formation. *Genes Dev.* **14**, 1377–1389 (2000).
31. Oberdick, J., Smeyne, R. J., Mann, J. R., Zackson, S. & Morgan, J. I. A promoter that drives transgene expression in cerebellar Purkinje and retinal bipolar neurons. *Science* **248**, 223–226 (1990).
32. Nagy, A., Gertsenstein, M., Vintersten, K. & Behringer, R. *Manipulating the Mouse Embryo: A Laboratory Manual*. (Cold Spring Harbor Lab., New York, 1986).
33. Song, Y. et al. Sample preparation project for the subcellular proteome of mouse liver. *Proteomics* **6**, 5269–5277 (2006).
34. Peng, F. et al. Proteomic and bioinformatics analyses of mouse liver microsomes. *Int. J. Proteomics* **2012**, 832569 (2012).
35. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **25**, 402–408 (2001).
36. Wolters, D. A., Washburn, M. P. & Yates, J. R. III An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683–5690 (2001).
37. McDonald, W. H. et al. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.* **18**, 2162–2168 (2004).
38. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50 (2003).
39. Boon, K. et al. Isolation and functional analysis of histidine-tagged elongation factor Tu. *Eur. J. Biochem.* **210**, 177–183 (1992).
40. Qin, D., Abdi, N. M. & Fredrick, K. Characterization of 16S rRNA mutations that decrease the fidelity of translation initiation. *RNA* **13**, 2348–2355 (2007).
41. Shoji, S., Walker, S. E. & Fredrick, K. Reverse translocation of tRNA in the ribosome. *Mol. Cell* **24**, 931–942 (2006).
42. Walker, S. E. & Fredrick, K. Preparation and evaluation of acylated tRNAs. *Methods* **44**, 81–86 (2008).
43. Merryman, C. & Green, R. Transformation of aminoacyl tRNAs for the in vitro selection of "drug-like" molecules. *Chem. Biol.* **11**, 575–582 (2004).
44. Langer, A. et al. Polymerase/DNA interactions and enzymatic activity: multi-parameter analysis with electro-switchable biosurfaces. *Sci. Rep.* **5**, 12066 (2015).
45. Splan, K. E., Musier-Forsyth, K., Boniecki, M. T. & Martinis, S. A. In vitro assays for the determination of aminoacyl-tRNA synthetase editing activity. *Methods* **44**, 119–128 (2008).
46. Beebe, K., Waas, W., Druzina, Z., Guo, M. & Schimmel, P. A universal plate format for increased throughput of assays that monitor multiple aminoacyl transfer RNA synthetase activities. *Anal. Biochem.* **368**, 111–121 (2007).
47. Moon, H. M. & Weissbach, H. Interaction of brain transferase I with guanosine nucleotides and aminoacyl-tRNA. *Biochem. Biophys. Res. Commun.* **46**, 254–262 (1972).
48. Sun, L. et al. Evolutionary gain of alanine mischarging to noncognate tRNAs with a G4:U69 base pair. *J. Am. Chem. Soc.* **138**, 12948–12955 (2016).
49. Naganuma, M. et al. The selective tRNA aminoacylation mechanism based on a single G·U pair. *Nature* **510**, 507–511 (2014).
50. He, L., Diedrich, J., Chu, Y. Y. & Yates, J. R. III Extracting accurate precursor information for tandem mass spectra by RawConverter. *Anal. Chem.* **87**, 11361–11367 (2015).
51. Xu, T. et al. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *J. Proteomics* **129**, 16–24 (2015).

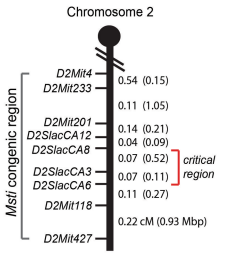
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

a

Strains	Affected *Aars* $^{sti/sti}$ mice
C57BL/6J	47/47
CAST/Ei **	22/46
MOLF/Ei *	22/22
DBA/2J *	33/33
BALB/cJ *	13/13
C3H/HeJ **	41/41
CASA/RkJ **	37/56

* F2 animals
** Backcross animals

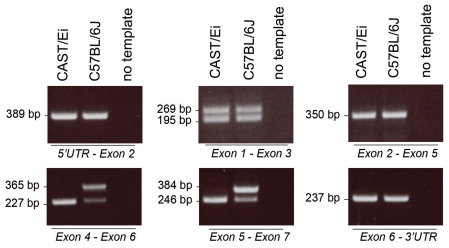
b



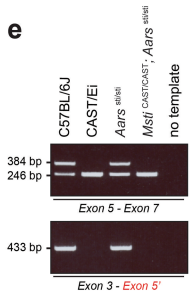
c

	Il2ra							*Fbxo18*			*Ankrd16*	
	T101S	Q106H	S143R	G158S	M225T	M233V	K236Q	H15Q	P369A	P376A	E335A	T346M
B6	-	-	-	-	-	-	-	-	-	-	-	-
MOLF	-	-	T -> A	G -> A	-	-	-	T -> G	C -> G	C -> G	A -> C	C -> T
DBA	-	-	-	-	-	-	-	-	-	-	-	-
BALB	-	-	-	-	-	-	-	-	-	-	-	-
C3H	-	-	-	-	-	-	-	-	-	-	-	-
CAST	C -> G	G -> T	T -> A	-	T -> C	A -> G	A -> C	T -> G	C -> G	C -> G	A -> C	C -> T
CASA	C -> G	G -> T	T -> A	-	T -> C	A -> G	A -> C	-	-	-	A -> C	C -> T

d



e

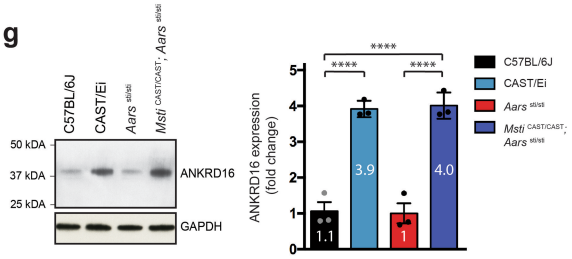


f

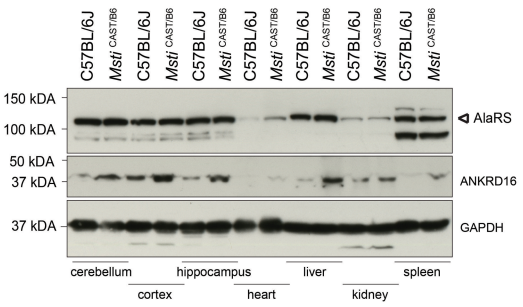
SNP detection - 5' splice site

Strain	SNP
C57BL/6J	GAGgta
DBA/2J	GAGgta
BALB/cJ	GAGgta
C3H/HeJ	GAGgta
MOLF/Ei	GAGgta
CAST/Ei	ggggta
CASA/RkJ	ggggta

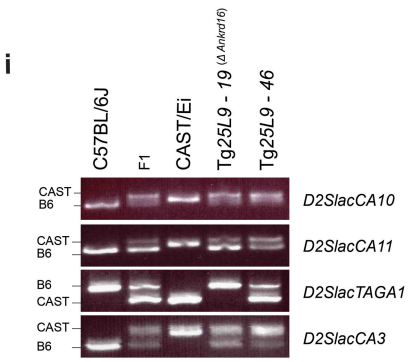
g



h



i



j

```
mouse       ----------MALPGDPRRLCRLVQEGRLRDLQEELAVARGC---------RGPAGDTLL  41
human       ----------MAQPGDPRRLCRLVQEGRLRALKEELQAAG---------GCPGPAGDTLL  41
rat         ----------MAPPGDPRRLCRLVQEGQLRALREELEVAGGCWDPEMFRGSQGPAGDTLL  50
chimp       ----------MAQLGDPRRLCRLVQEGRLRALKEELQAAG---------GSPGPAGDTLL  41
cattle      ----------MAAPADPRRLCRLVQDGRLCTLREELQAAGH--------AGCAGPAGDTLL  43
chicken     ----------MAEWPRLLTLARQGKLDLLEAALR-------DGDVTNRRCDRSGDTLL  41
zebrafish   ----------MEEENTVKLLIKAVQHGELCTVKQHLNELADN-KQTVIYTHFGKSGDTVL  49
frog        MLNDNSSLNKMNEEDQYKHMLRIIQEGQLGLLQEQLQREPHLL-GALNVKHFGKSGDTLL  59

mouse       HCAARHGRQDILAYLVEAWSMDIEATNRDYKRPLHEAASMGHRDCVRYLLGRGAVVDSLK 101
human       HCAARHGHRDVLAYLAEAWGMDIEATNRDYKRPLHEAASMGHRDCVRYLLGRGAAVDCLK 101
rat         HFASRHGRQDILAYLVEAWSMDIEAANRDYKRPLHEAASMGHRDCVRYLLGRGAVVDSLK 110
chimp       HCAARHGHRDVLAYLAEAWGMDIEATNRDYKRPLHEAASMGHRDCVRYLLGRGAAVDCLK 101
cattle      HCAARHGRRDVLAYLVEAWDLDIEAANRDYKRPLHEAASMGHRNCVRYLLGRRAAVDCLK 103
chicken     HVAARHGHLHLLVLLLETLGMDVEVANGDYKRPLHEAASMGHRECVSYLLAKGASVDCLK 101
zebrafish   HYAARHGHVDILRYLVEELNMDIEVYNSDYKRALHEASSMGHYECVRFLICRGAKVDSLK 109
frog        HYAARHGHLSILTYLVEAVGMDVEVLNNDYKRPLHEAASMGHRDCLLYLLSRGTEVDCLK 119

mouse       KADWTPLMMACTRKNLDVIQDLVEHGANPLLKNKDGWNSFHIASREGHPVILRYLLTVCP 161
human       KADWTPLMMACTRKNLGVIQELVEHGANPLLKNKDGWNSFHIASREGDPLILQYLLTVCP 161
rat         KADWTPLMMACTRKNLEVIQDLVEHGANPLLKNKDGWNSFHIASREGDPVILRYLLTVCP 170
chimp       KADWTPLMMACTRKNLGVIQELVEHGANPLLKNKDGWNSFHIASREGDPLILQYLLTVCP 161
cattle      KADWTPLMMACTRRNLEVIRDLVEHGANPLLKNKDGWNSFHIASREGDPLILQYLLAVCP 163
chicken     RADWTPLMMACTRQNLEVIKELVEHGANPLLKNKDGWNCFHIASREGHLEVLQYLLAVFP 161
zebrafish   KADWTPLMMACTRRNLEVIQELLDYGADPTLQNKDGWSSFHIACREGDPAIINHLLLAKP 169
frog        KADWTPLMMACTKTKLDIIKDLIEHRANPMLKNKDGWNCFHIATREGDVAIIQYLLDVFP 179

mouse       DAWKTESNIRRTPLHTAAMHGCLEAVQVLLERCHYEPDCRDNCGVTPFMDAIQCGHVSIA 221
human       GAWKTESKIRRTPLHTAAMHGHLEAVKVLLKRCQYEPDYRDNCGVTALMDAIQCGHIDVA 221
rat         DVWKTESKIRRTPLHTAAMHGCFEAVQELLERCHYEPDCRDNCGVTPFMDAIQCGHVSIA 230
chimp       GAWKTESKIRRTPLHTAAMHGHLEAVKVLLKRCQYEPDYRDNCGVTALMDAIQCGHIDVA 221
cattle      SAWRTESKIGRTPLHTAAMHGCLEAVEVLLQRCQYEPDCRDRCGLTPFMDAVQCGHTDVA 223
chicken     SSWDTESTIRRTPLHTAAMHGCFDAVGLLLERCHYKPDSRDNCGVTPFMDAIQNGHVNIA 221
zebrafish   DIWKTKSKTGRTPLHTAAMHGCEDAVKILLQSCCYEPDEKDSCGVTPFMDAIRNGHIAVA 229
frog        DIWKTESKIKRTPLHTAAMHGCFEVIEVLLERCNYDPDCKDSCGVTPFMDAVQNGHLSIA 239

mouse       KLLLEQHKACSSAADSMGAQALHRAAVTGQDEAIRFLVCGLGIDVDVRAKSSQLTALHYA 281
human       RLLLDEHGACLSAEDSLGAQALHRAAVTGQDEAIRFLVSELGVDVDVRATSTHLTALHYA 281
rat         KLLLENHEACSSATDSLGAQALHRAAVTGQDEAIRFLVCGLGIDVDVRAKSSQLTALLYA 290
chimp       RLLLDEHGACLSAEDSLGAQALHRAAVTGQDEAIRFLVSELGVDVDVRATSTHLTALHYA 281
cattle      RLLLKKHKACVSAEDSLGAQAIHRAAVTGQDEAIRFLVSELGVDVDARAGSTHLTALHFA 283
chicken     RLLLEKHQACPTAVDALGAQPLHRAAVTAQDEAIQFLVSELGVDVNGRATPLQLTALHYA 281
zebrafish   KLLLEDHQASPSAVDILGAQPLHQASVTAQEEALSFLVHHLGMDINSRATKLELTALHYA 289
frog        QLLIEKKKVCYSAFDRMGAQALHLAAVTGQNESLQYLVSHLGVNVNEKATSMQLSALHYA 299

mouse       AKEGQTNTVQTLLSLGADINSTDERNRSVLHLACAGQHVACTRLLLQSGLKDSEDLTGTL 341
human       AKEGHTSTIQTLLSLGADINSKDEKNRSALHLACAGQHLACAKFLLQSGLKDSEDITGTL 341
rat         AKEGQTSTVQTLLSLGADINSTDERNRSALHLACAGQHAACARFLRQSGLKDSADLTGAL 350
chimp       AKEGHTSTIQTLLSLGADINSKDEKNRSALHLACAGQHLACAKFLLQSGLKDSEDITGTL 341
cattle      AKEGHVSTVQTLLSLGADINSKDERNRSALHLACAGQHAACVQFLLSSGLQDSPDVTGTL 343
chicken     AKEGHMCTIQTLLSLGANIHAKDGKGRSALHAACAGQHAAAAQILLSRGLQDSEDSTGTL 341
zebrafish   AKDGHTSTIKTLLALGADLHARDAKGRSALHMACIGQHVDTVRMLLQMGLTDTEDNTGNT 349
frog        AKEGHVGTVLTLLTLGADLHCNDLKGRSALHMASASQHASCVHILLKSGLHDSADCTGVY 359

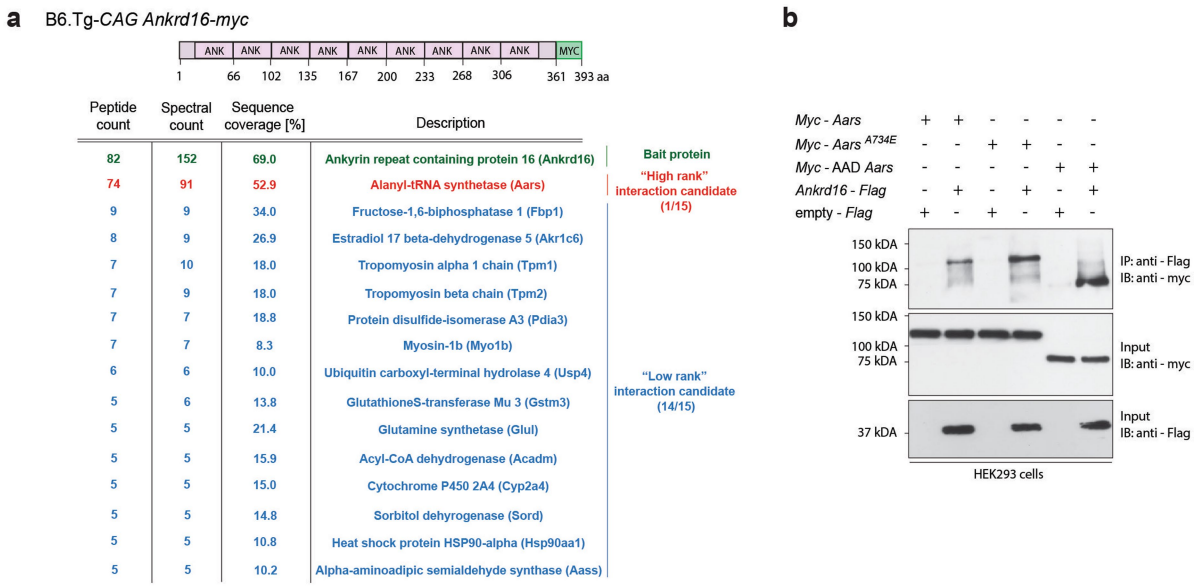
mouse        AQQERSVDILQDFDHDVKS--------                                361
human       AQQLPRRADVLQGSGHSAMT--------                                361
rat         AQQLTRGVAILQDFDPGVKS--------                                370
chimp       AQQLPRRADVLQGSGHSAMT--------                                361
cattle      AQHLTRSPNVLQCFNHSTMAEGISRDSQ                                371
chicken     AQQLARKPDVIQVFQRTNVNS-------                                362
zebrafish   ARQYAIKPHIKDVFENLAT---------                                368
frog        AKDLAKKPEVHQVFKNIC---------                                377
```

Extended Data Fig. 1 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 1 | Identification of *Ankrd16* as a genetic modifier of the B6J.*Aars*$^{sti/sti}$ mutation. a, Three- to four-month-old *Aars*$^{sti/sti}$ mice from crosses to inbred strains, and the numbers affected with ataxia and Purkinje cell degeneration. **b**, Location of *Msti* relative to microsatellite markers. cM, centimorgan. **c**, Non-synonymous SNPs in *Msti* candidate genes. The amino acids and positions are shown at the top of the table with the B6 residue listed first. **d**, RT–PCR analysis of *Ankrd16* transcripts from cerebellar cDNA prepared from C57BL/6J and CAST/Ei mice. Note, the alternative *Ankrd16* transcript containing exon 5′ (see Fig. 2c) amplified by primers to exon 4/6 or exon 5/7 was present in cDNA from C57BL/6J but not CAST/Ei mice, whereas the alternative transcript detected by exon1/3 primers was present in cDNA from both strains. $n = 2$ biological replicates. **e**, RT–PCR analysis of *Ankrd16* transcripts from cerebellar cDNA prepared from C57BL/6J, CAST/Ei, *Aars*$^{sti/sti}$ and *Msti*$^{CAST/CAST}$*Aars*$^{sti/sti}$ mice. Note, the alternative *Ankrd16* transcript containing exon 5′ (see Fig. 2c) amplified by exon 5/7 or 3/5′ primers was present in cDNA from C57BL/6J and *Aars*$^{sti/sti}$, but absent in the presence of CAST/Ei-derived *Msti* (CAST/Ei and *Msti*$^{CAST/CAST}$*Aars*$^{sti/sti}$ mice). $n = 3$ biological replicates. **f**, Sequence of the SNP-containing region in intron 5 of *Ankrd16*. Upper-case letters indicate novel exon 5′ and lower-case letters indicate intron 5. The SNP in non-rescuing strains is shown in red. **g**, Western blotting analysis of ANKRD16 from cerebellar lysates. Note that the expression of ANKRD16 is reduced in C57BL/6J and *Aars*$^{sti/sti}$ mice relative to mice with CAST/Ei-derived *Msti* (CAST/Ei and *Msti*$^{CAST/CAST}$*Aars*$^{sti/sti}$ mice) (mean ± s.d., $n = 3$, one-way ANOVA (Tukey correction), $^{****}P \leq 0.0001$). **h**, Protein levels of AlaRS and ANKRD16 were determined by western blotting using various mouse tissues from C57BL/6J and B6 congenic mice heterozygous for the *Msti* region (*Msti*$^{CAST/B6}$). GAPDH is included as a loading control. Note the increase in ANKRD16 levels in *Msti*$^{CAST/B6}$ tissues, whereas AlaRS levels do not change between genotypes. $n = 2$ biological replicates for all tissues, $n = 4$ for neuronal tissues. **i**, PCR results of genomic DNA from B6 transgenic mouse lines Tg25L9-19 and Tg25L9-46, which carry the CAST/Ei BAC. Polymorphic markers, which differentiate between C57BL/6J and CAST/Ei were used as shown. $n = 5$ biological replicates. **j**, Amino acid sequence comparison of ANKRD16 from various species with the C57BL/6J strain shown. Non-synonymous SNPs distinguishing CAST/Ei and C57BL/6J are shown in yellow and serinylated lysines are shown in red.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



f

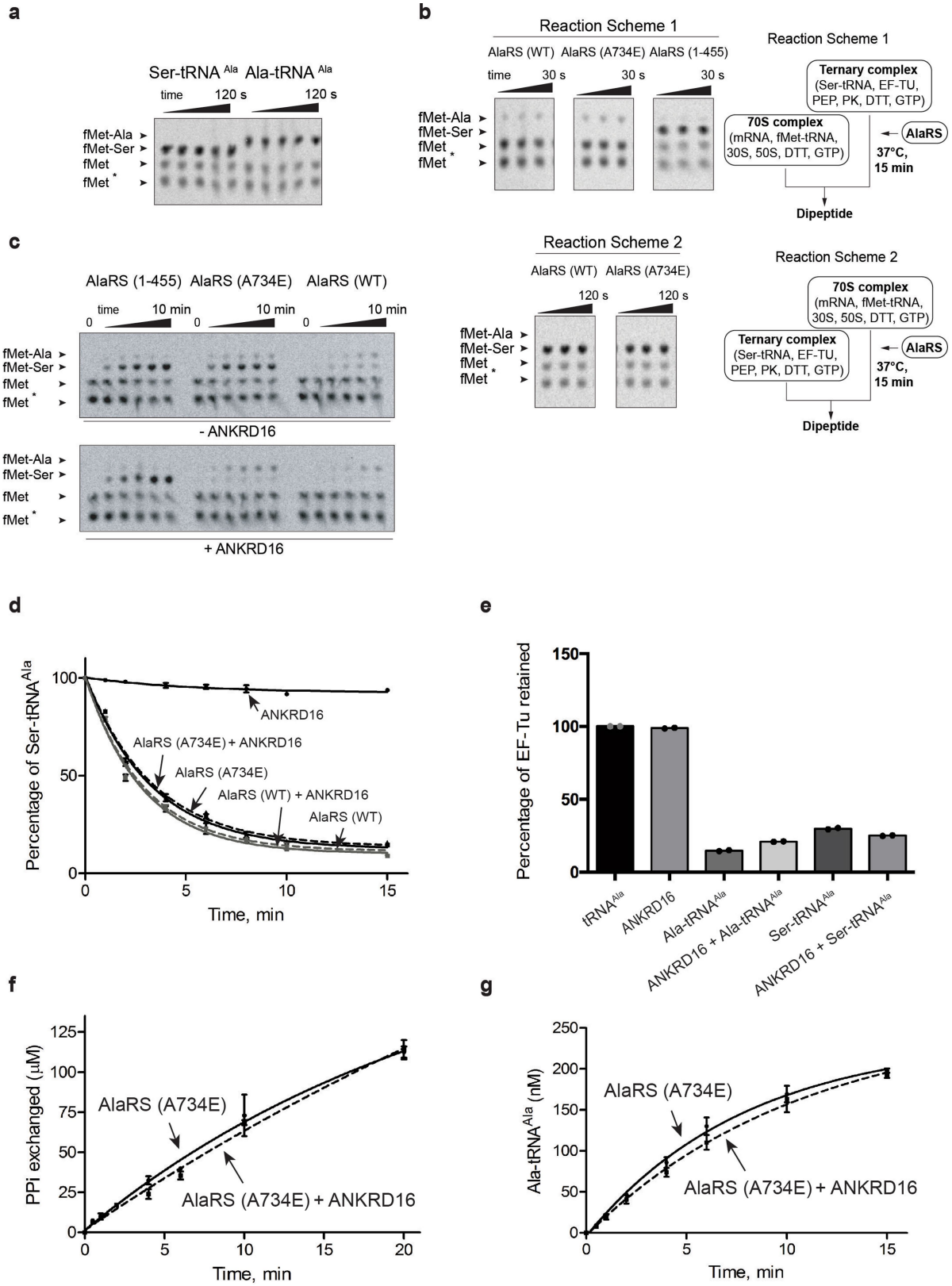
	k_{ON} (M^{-1} s^{-1})	k_{OFF} (s^{-1})	K_D (μM)
AlaRS / ANKRD16	5310 ± 240	0.0105 ± 0.0001	1.98 ± 0.09
AlaRS (A734E) / ANKRD16	3985 ± 180	0.0076 ± 0.0003	1.90 ± 0.12
AlaRS (1-455) / ANKRD16	3660 ± 140	0.0087 ± 0.0002	2.37 ± 0.10
AlaRS (A734E) / ANKRD16 (3xR)	1915 ± 160	0.0103 ± 0.0002	5.41 ± 0.46

Extended Data Fig. 2 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 2 | Verification of the interaction between ANKRD16 and AlaRS in vitro. a, Peptide/spectral counts of proteins co-immunoprecipitated from transgenic *Ankrd16*-Myc (see diagram) but not detected in non-transgenic liver tissue. $n = 1$ experiment. **b**, HEK293T cells were transiently co-transfected with Myc-tagged constructs for mouse *Aars*, *Aars*^{A734E}, the *Aars* aminoacylation domain (AAD) and Flag-tagged *Ankrd16*. Co-IP experiments were performed with ANKRD16–Flag as the bait protein. $n = 3$ independent experiments. **c**, Reciprocal co-IP experiments were performed by transiently co-transfecting HEK293T cells with HA-tagged constructs for mouse *Aars* or *Aars*^{A734E} and the Flag-tagged *Ankrd16*. HA-AlaRS proteins were used as bait for pull down. $n = 3$ independent experiments. **d**, Domain structure of mouse ANKRD16 and ANKRD29. HEK293T cells were transiently transfected with Flag-tagged constructs for mouse *Ankrd16*, *Ankrd29* and the HA-epitope tagged *Aars*. Co-IP experiments were performed with HA-AlaRS as bait protein. $n = 2$ independent experiments. **e**, Various domain protein products of AlaRS (human) as indicated were bacterially expressed, purified, and incubated with GST–ANKRD16. GST-pull down products and input were immunoblotted with anti-His or anti-GST antibodies. Asterisks indicate protein degradation products. $n = 2$ independent experiments. **f**, Binding dynamics were determined between mouse wild-type or mutant AlaRS and mouse wild-type or mutant ANKRD16 using switchSENSE (mean ± s.d.; $n = 3$).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

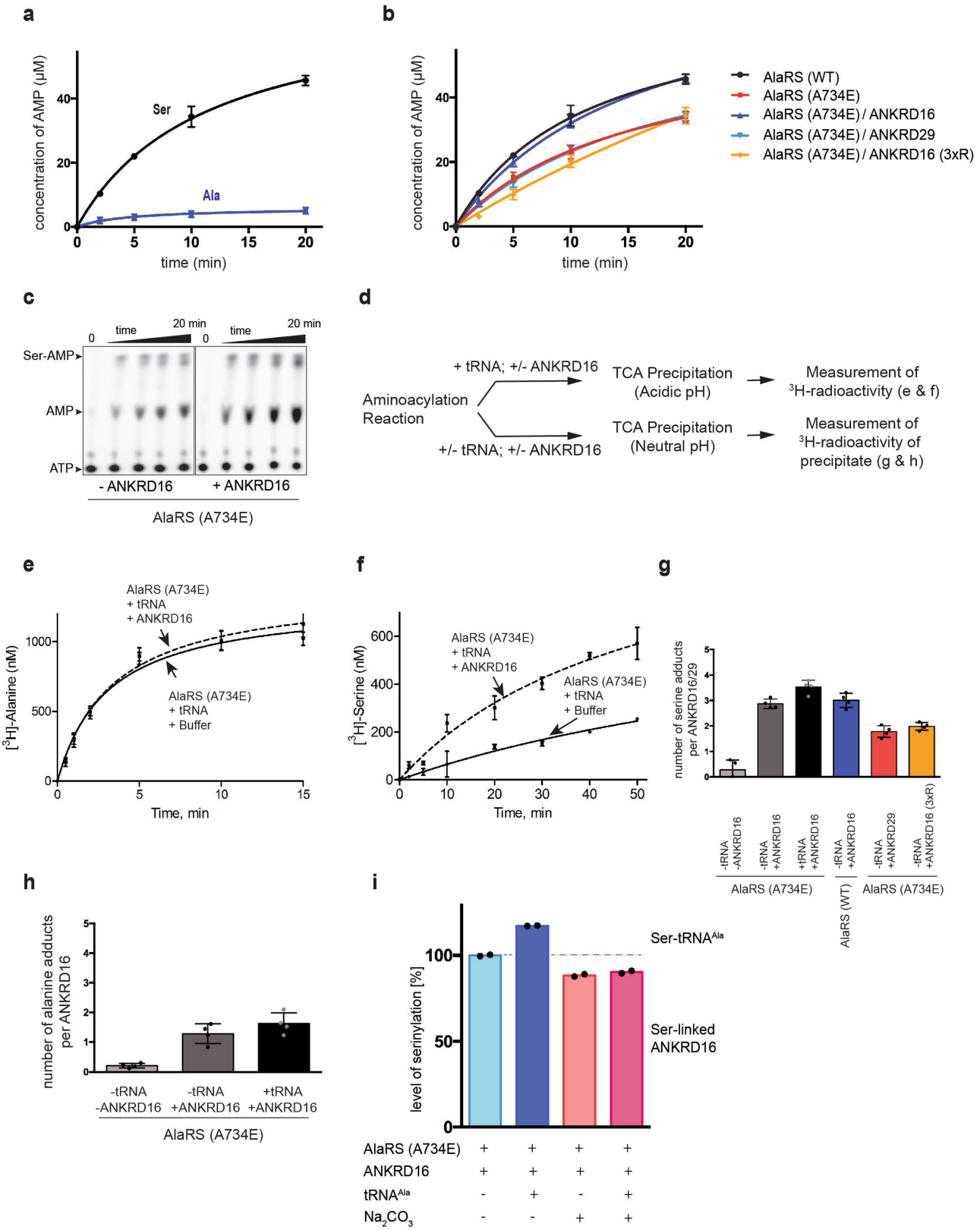


Extended Data Fig. 3 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 3 | Analysis of the effects of ANKRD16 on steps of translation. a, Preassembled ternary complex containing either Ser-tRNA^{Ala} or Ala-tRNA^{Ala} was mixed with 70S initiation complex programmed with codon GCU in the A site, aliquots were transferred at various times to a quenching solution (0.5 M KOH), and products were resolved by eTLC. $n = 1$ experiment. **b**, Incubation of full-length AlaRS with preassembled EF-Tu·GTP·Ser-tRNA^{Ala} prevents fMet-Ser formation. EF-Tu·GTP·Ser-tRNA^{Ala} ternary complex and 70S initiation complex were each preassembled. In reaction scheme 1, AlaRS was incubated with EF-Tu·GTP·Ser-tRNA^{Ala} for 15 min, 70S initiation complex was added, and aliquots were removed at various times for eTLC analysis. By contrast, in reaction scheme 2, AlaRS was incubated with the 70S initiation complex for 15 min, followed by addition of EF-Tu·GTP·Ser-tRNA^{Ala}, and aliquots were removed at various times for eTLC analysis. $n = 2$ independent experiments. **c**, Deacylated tRNA^{Ala} was mixed with AlaRS, serine, ATP and all other components to form the ternary complex, aliquots were transferred at various time points to the 70S initiation complex, and dipeptide products were resolved by eTLC. $t = 0$ indicates control reactions in the absence of AlaRS; * indicates oxidized fMet. Note, without alanine supplementation (**b** and **c**), trace amounts of fMet-Ala are detected, probably due to AlaRS-bound alanyl-AMP during protein purification. $n = 2$ independent experiments. **d**, Deacylation of [^{3}H]Ser-tRNA^{Ala} by mouse wild-type AlaRS or AlaRS(A734E) in the presence or in the absence of ANKRD16 (mean ± s.d., $n = 2$, one-phase decay model. R^2 values are as follows: wild-type AlaRS, 0.9892; AlaRS(A734E), 0.991; wild-type AlaRS + ANKRD16 = 0.9872; AlaRS(A734E) + ANKRD16, 0.9902; ANKRD16, 0.7992). **e**, Percentage of EF-Tu retained on the filter membrane upon the addition of various components as indicated (mean ± s.d.; $n = 2$). **f**, ATP-pyrophosphate exchange by mouse AlaRS(A734E) in the presence or in the absence of ANKRD16 (mean ± s.d., $n = 2$, Michaelis–Menten model. R^2 values: AlaRS(A734E), 0.9763; AlaRS(A734E) + ANKRD16, 0.9874. **g**, Aminoacylation of tRNA^{Ala} with alanine by mouse AlaRS(A734E) in the presence or in the absence of ANKRD16 (mean ± s.d., $n = 2$, Michaelis–Menten model. R^2 values: AlaRS(A734E), 0.9879; AlaRS(A734E) + ANKRD16, 0.9846.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

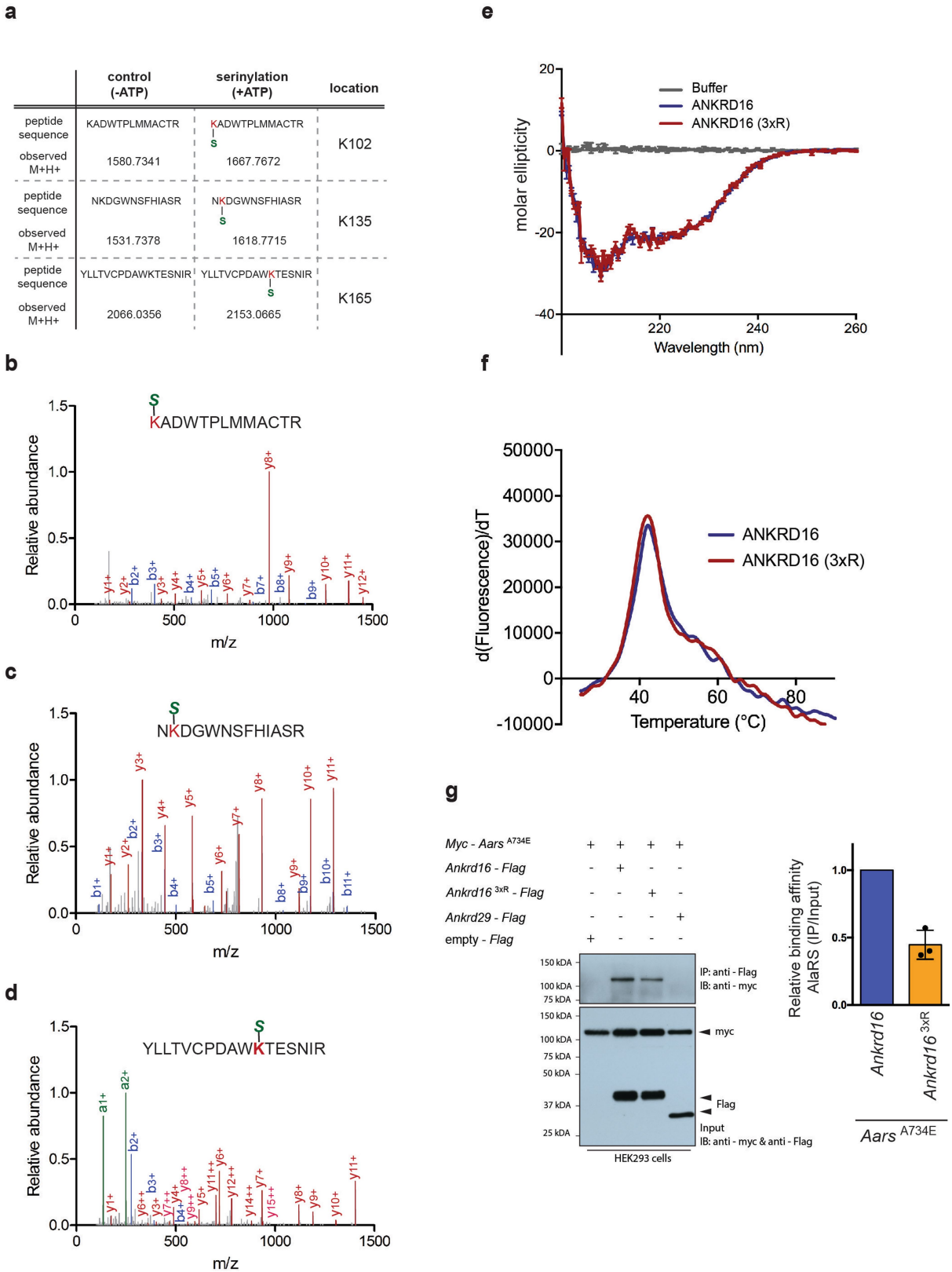


Extended Data Fig. 4 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 4 | Analysis of serinylation of ANKRD16. a, tRNA-independent ATPase activity of mouse wild-type AlaRS for serine or alanine (mean ± s.d., $n = 3$, Michaelis–Menten model. R^2 for serine, 0.9926; for alanine, 0.8430. **b**, tRNA-independent ATPase activity of mouse wild-type AlaRS or AlaRS(A734E) for serine in the presence of ANKRD16, ANKRD29, or ANKRD16(3×R) (mean ± s.d., Michaelis–Menten model. R^2 values are as follows: wild-type AlaRS, 0.9926; AlaRS(A734E), 0.9899; AlaRS(A734E) + ANKRD16, 0.9918; AlaRS(A734E) + ANKRD29, 0.9939; AlaRS(A734E) + ANKRD16(3×R), 0.9841. For wild-type AlaRS, AlaRS(A734E) and AlaRS(A734E) + ANKRD29, $n = 3$; for AlaRS(A734E) + ANKRD16 and AlaRS(A734E) + ANKRD16(3×R), $n = 4$. **c**, TLC analysis of ATPase activity of AlaRS(A734E) against serine in the absence and in the presence of ANKRD16. $n = 3$ independent experiments. **d**, Experimental scheme showing how data were generated for **e**, **f**, **g** and **h**. **e**, Acylation reactions with radioactive alanine were performed as described above to determine alanine transfer (mean ± s.d., $n = 2$, Michaelis–Menten model. R^2 values: AlaRS(A734E) + tRNA + ANKRD16, 0.9611; AlaRS(A734E) + tRNA + buffer, 0.9641. **f**, Misacylation of $tRNA^{Ala}$ with radioactive serine by mouse AlaRS(A734E) in the presence or in the absence (buffer) of ANKRD16. After misacylation, reactions were subjected to TCA precipitation of RNA (Ser-$tRNA^{Ala}$) and protein under acidic conditions to maintain Ser-tRNA (mean ± s.d., $n = 2$, Michaelis–Menten model. R^2 values AlaRS(A734E) + tRNA + ANKRD16, 0.9577; AlaRS(A734E) + tRNA + buffer, 0.7841. **g**, Misacylation reactions using radioactive serine and mouse wild-type AlaRS or AlaRS(A734E) were performed in the presence or in the absence of ANKRD16, ANKRD29, ANKRD16(3×R) or $tRNA^{Ala}$. After TCA precipitation under neutral pH conditions, serine was measured (mean ± s.d., $n = 4$). Note the higher level of TCA-precipitated serine on tRNA or protein in the presence of ANKRD16. **h**, Acylation reactions using radioactive alanine and mouse AlaRS(A734E) were performed in the presence or in the absence of ANKRD16 or $tRNA^{Ala}$. After TCA precipitation under neutral pH conditions, alanine was measured (mean ± s.d., $n = 4$). **i**, Misacylation reactions using radioactive serine and mouse AlaRS(A734E) were performed in the presence of ANKRD16 either with or without $tRNA^{Ala}$. Reactions were treated with or without Na_2CO_3 (final concentration of 0.15 M, alkaline pH) for 30 min, followed by TCA precipitation. Precipitated [^{3}H]serine-links were measured and plotted as relative level of serinylation (mean, $n = 2$).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

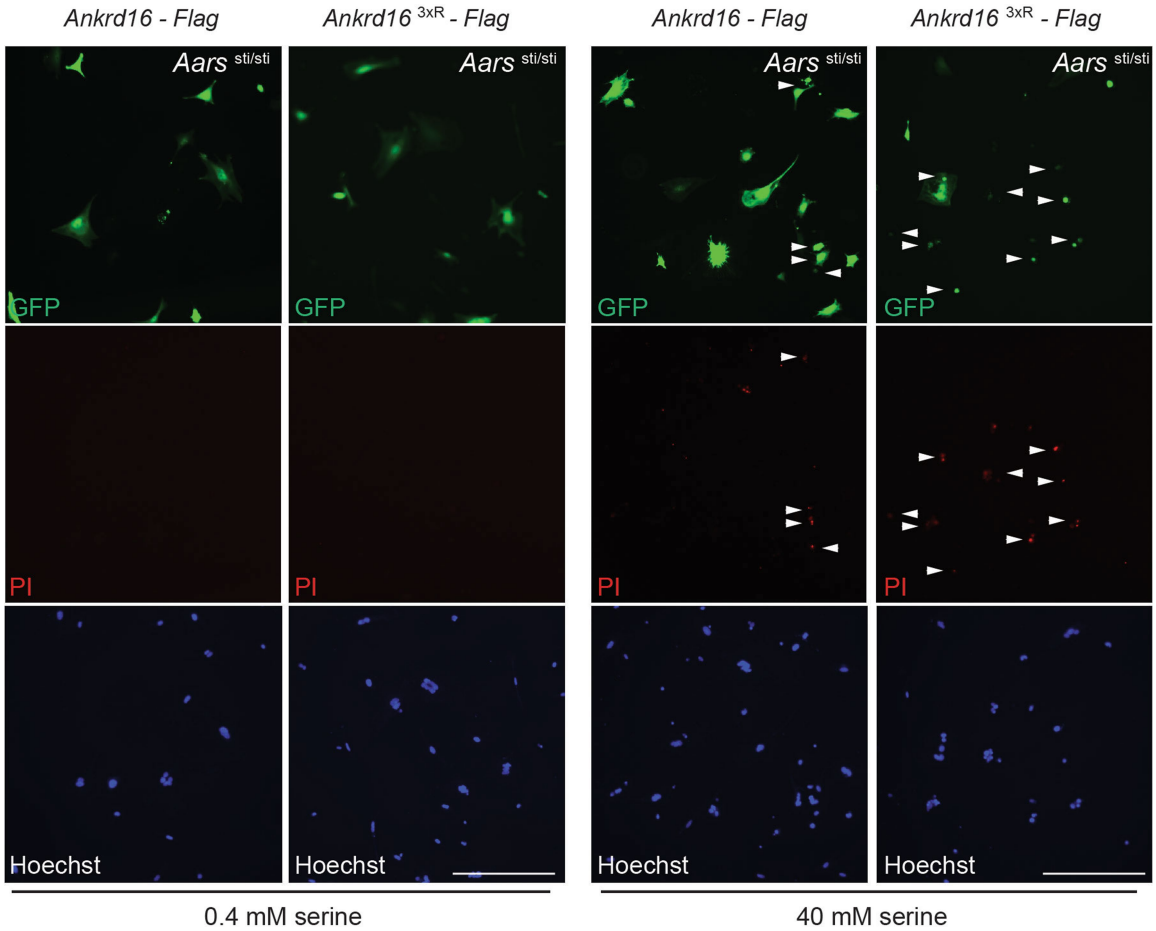


Extended Data Fig. 5 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

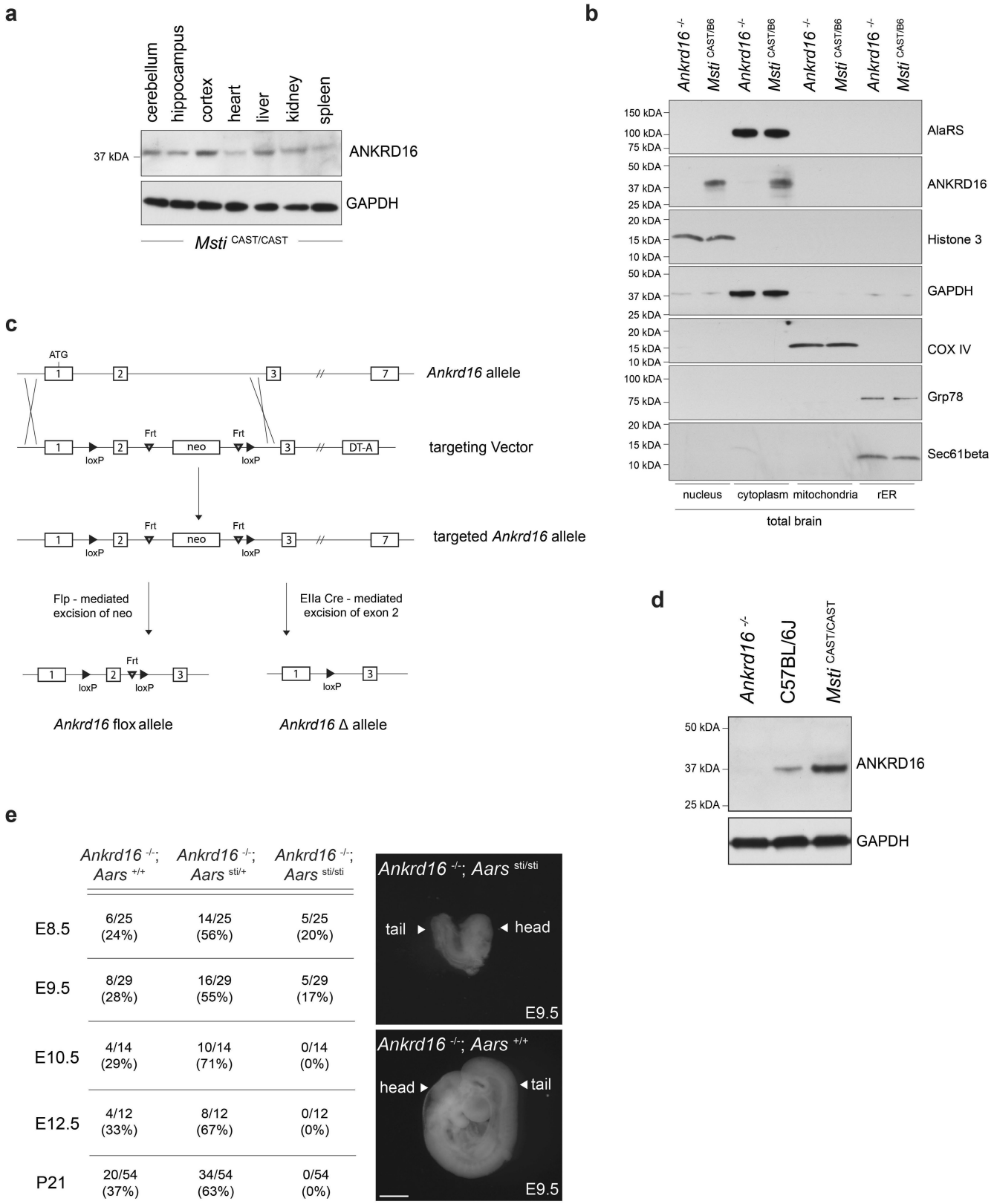
Extended Data Fig. 5 | Analysis of serinylation of ANKRD16 by mass spectrometry. a–**d**, MS/MS spectra of peptides from ANKRD16. Incorporation of serine onto ANKRD16 was observed when serine was misactivated by mouse AlaRS(A734E) (+ATP). Misactivation was not observed in the absence of ATP. MS/MS spectrum of peptides from ANKRD16 with serine linked to positions K102 (**b**), K135 (**c**) and K165 (**d**). a ions (**a**), b ions (**b**) and y ions (y) are annotated in green, red and purple, respectively. The triply charged precursor had a mass of 2,153.065 daltons and included carbamidomethyl cysteine. For **a**–**d**, $n = 1$ experiment. **e**, Secondary structure analysis of mouse ANKRD16 and ANKRD16(3×R). Far-UV circular dichroism spectra of wild-type ANKRD16 (blue) and ANKRD16(3×R) (red) show highly similar CD spectra (mean ± s.d., $n = 4$). **f**, Thermal-shift analysis of mouse ANKRD16 (blue) and ANKRD16(3×R) (red) show highly similar thermal stability. $n = 2$ independent experiments. **g**, HEK293T cells were transiently co-transfected with Myc-tagged constructs for mouse AlaRS(A734E) and Flag-tagged ANKRD16, ANKRD16(3×R) or ANKRD29. Co-IP experiments were performed with Flag-tagged proteins as the bait protein. Binding affinity was determined by normalizing the immunoprecipitation signal of AlaRS(A734E) to the input signal, in which the interaction of AlaRS(A734E) + ANKRD16 was arbitrarily defined as 1 to determine the relative binding affinity of AlaRS(A734E) + ANKRD16(3×R) (mean ± s.d., $n = 3$).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Serine-induced cell death in *Aars*$^{sti/sti}$ fibroblasts. B6.*Aars*$^{sti/sti}$ embryonic fibroblasts were co-transfected with hrGFP (humanized recombinant GFP, green) and either ANKRD16–Flag or ANKRD16(3×R)–Flag ($n = 4$). 12 h post-transfection, serine was added and cells were cultured for 24 h before staining with propidium iodide (PI; red) and Hoechst (blue) to determine cell death. Arrowheads represent PI^+GFP^+ cells. $n = 4$ biological replicates. Scale bars, 100 μm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

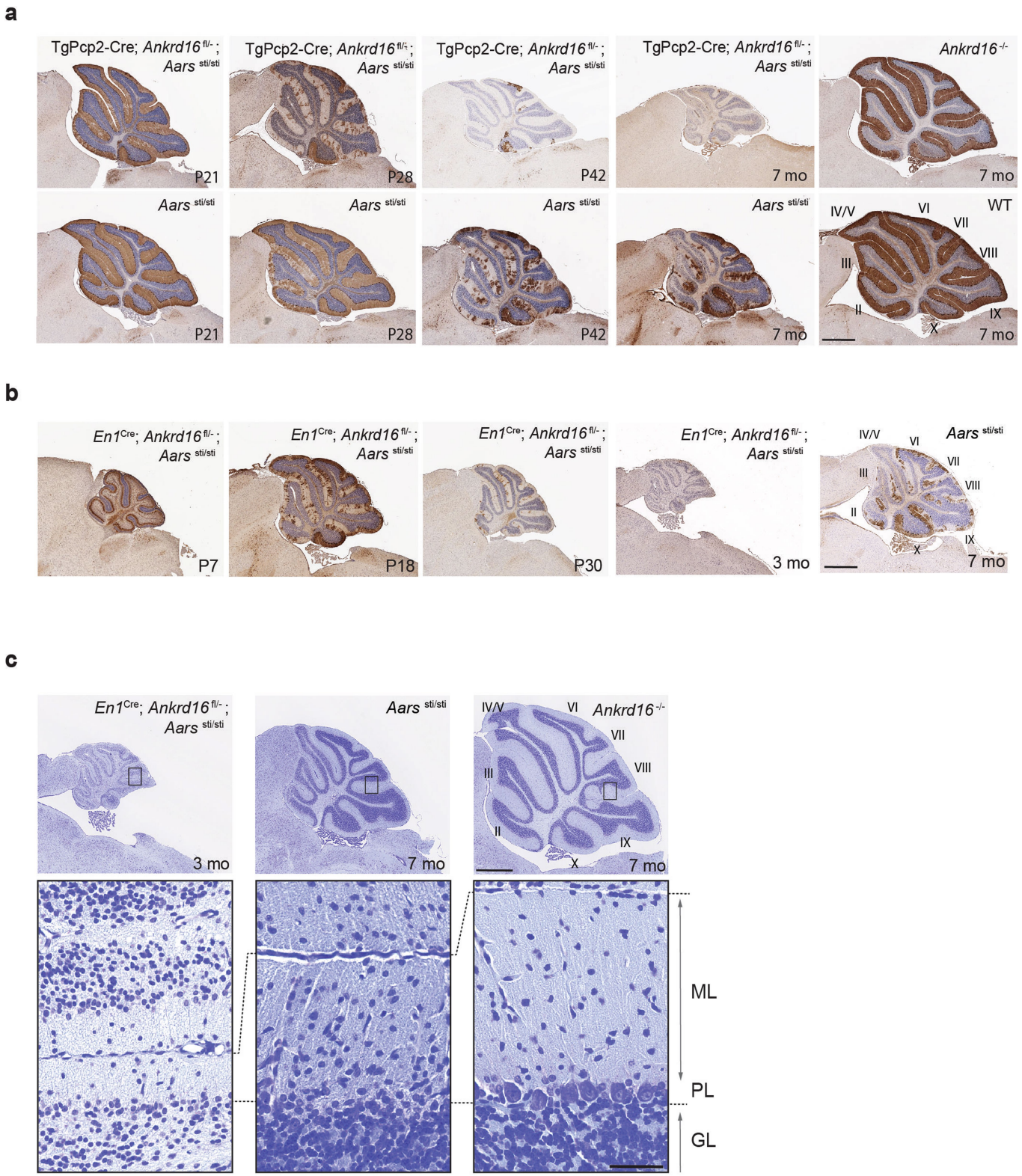


Extended Data Fig. 7 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

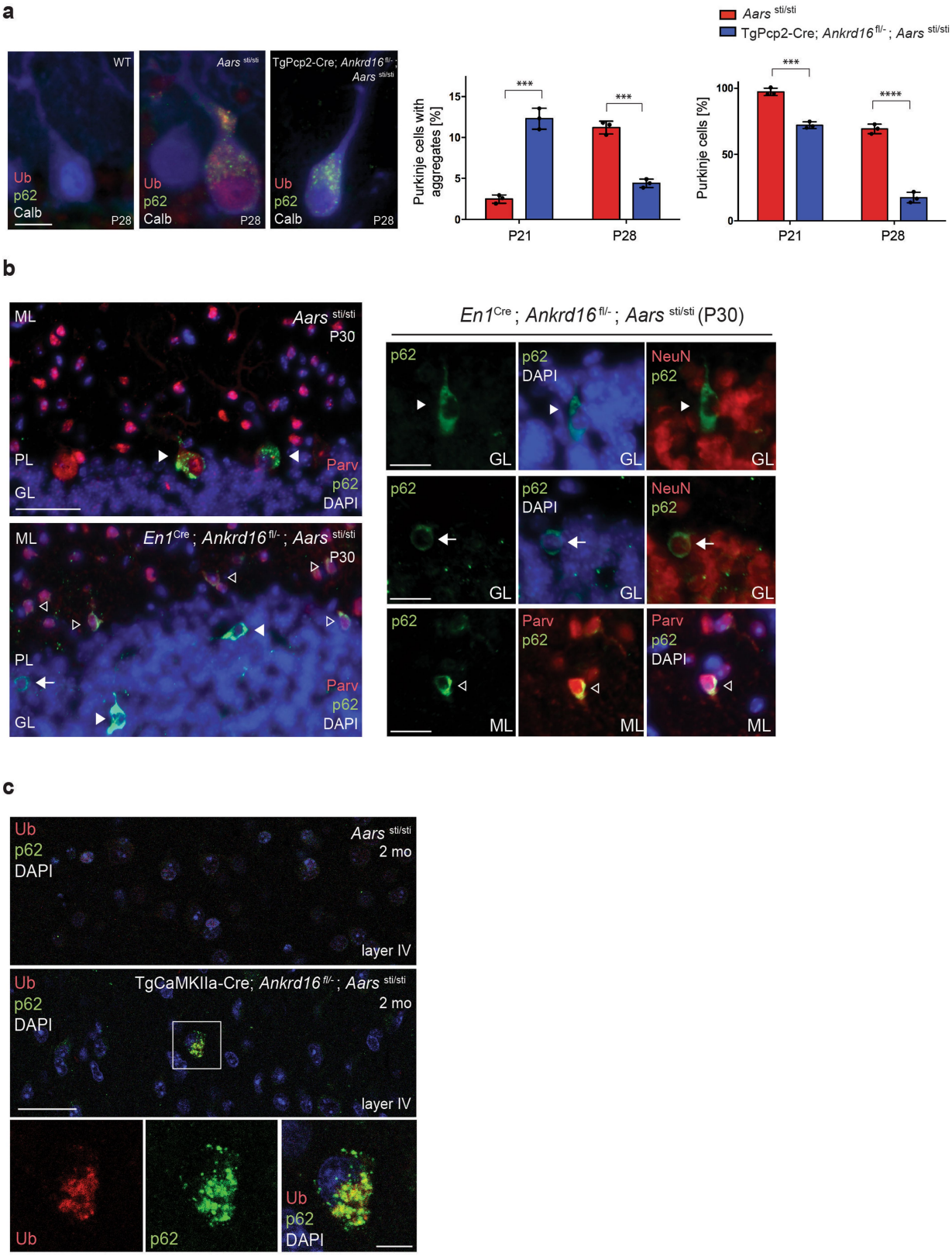
Extended Data Fig. 7 | Analysis of ubiquitous deletion of ANKRD16. **a**, Protein levels of ANKRD16 were determined by western blotting of tissues from B6 mice homozygous for the *Msti* region ($Msti^{CAST/CAST}$). $n = 3$ biological replicates. **b**, Subcellular analysis of ANKRD16 and AlaRS by cell fractionation of brains from B6.$Msti^{CAST/B6}$ and $Ankrd16^{-/-}$ mice. Cellular fractions were confirmed using antibodies for histone 3 (nuclear marker), GAPDH (cytosolic marker), COX IV (mitochondrial marker), GRP78 (endoplasmic reticulum marker), and Sec61 beta (rough endoplasmic reticulum marker). $n = 2$ independent experiments. **c**, To generate a ubiquitous or conditional loss-of-function allele, *lox*P sites that flank exon 2 of *Ankrd16* were inserted by homologous recombination. Removal of exon 2 results in a frame shift and premature stop codon in exon 3. *Ankrd16* was ubiquitously deleted by EIIa Cre-mediated removal of exon 2 and the neo cassette. Flippase (Flp)-mediated excision of the neo cassette was used to generate a conditional loss-of-function *Ankrd16* allele. **d**, Loss of ANKRD16 was verified by western blotting. Protein extracts from $Ankrd16^{-/-}$, $Msti^{CAST/CAST}$ and C57BL/6J mice were used for comparison purposes and GAPDH was used as a loading control. $n = 4$ biological replicates. **e**, The number of embryos or mice of various genotypes from intercrosses of $Aars^{sti/+}Ankrd16^{-/-}$ mice over the total number observed. Representative images of E9.5 embryos. Note, $Aars^{sti/sti}Ankrd16^{-/-}$ embryos are smaller and have failed to turn. E, embryonic day; P, postnatal day. For E8.5 and E9.5, $n = 4$ litters for each time point; E10.5 and E12.5, $n = 2$ litters for each time point; P21, $n = 7$ litters. Scale bars, 500 μm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 8 | Conditional deletion of *Ankrd16* accelerates Purkinje cell loss and causes widespread neurodegeneration in the B6.*Aars*$^{sti/sti}$ cerebellum. a, **b**, Calbindin D-28 immunohistochemistry of sagittal cerebellar sections. **c**, Cresyl violet-stained sagittal cerebellar sections. Note the presence of interneurons in the molecular layer (ML) of B6.*Aars*$^{sti/sti}$ cerebellum despite the thinning of the molecular layer as a consequence of Purkinje cell degeneration. By contrast, loss of *Ankrd16* in the B6.*Aars*$^{sti/sti}$ cerebellum results in degeneration of molecular-layer interneurons. GL, granule cell layer; PL, Purkinje cell layer. The number of biological replicates are as follows: for TgPcp2-Cre *Ankrd16*$^{fl/-}$*Aars*$^{sti/sti}$ or *En1*^{cre}*Ankrd16*$^{fl/-}$*Aars*$^{sti/sti}$: P7, $n=5$; P18, $n=4$; P21, $n=4$; P28, $n=7$; P30, $n=5$; P42, $n=4$; 3 months, $n=5$; 7 months, $n=6$. For *Ankrd16*$^{-/-}$: 7 months, $n=7$. For *Aars*$^{sti/sti}$ and wild-type: 7 months, $n=3$. Scale bars, 500 μm (**a**, **b**, **c**), 50 μm (**c**, higher magnification).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

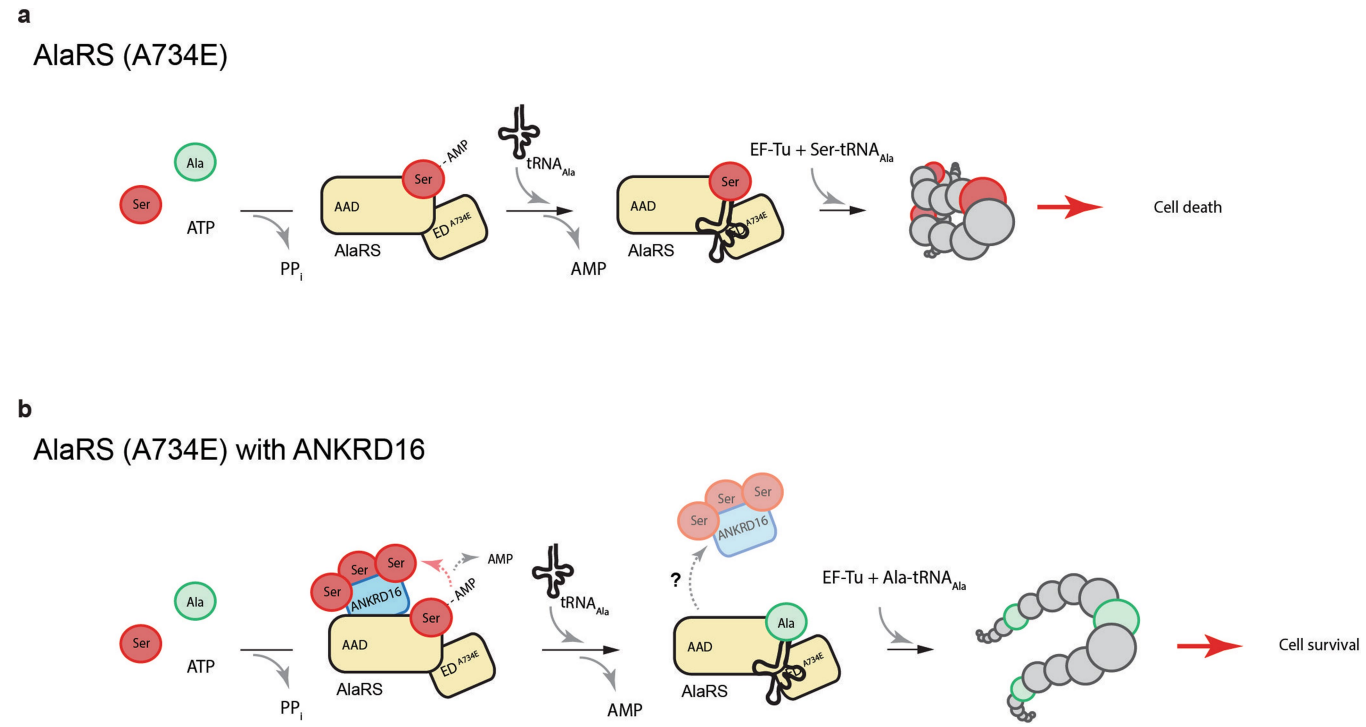


Extended Data Fig. 9 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 9 | Conditional deletion of *Ankrd16* accelerates the formation of protein aggregates in B6.*Aars*$^{sti/sti}$ mice. a, Ubiquitin (Ub; red), p62 (green), and Calbindin D-28 (Calb; blue) immunofluorescence on sagittal cerebellar sections. The percentage of aggregate-positive Purkinje cells and Purkinje cells are shown (mean ± s.d., $n = 3$, multiple *t*-tests (Holm–Sidak method), *** $P = 0.0002466$ (Purkinje cells with aggregates, P21), $P = 0.0002336$ (Purkinje cells with aggregates, P28), $P = 0.0003214$ (percentage of Purkinje cells, P21), **** $P = 7.701787 \times 10^{-5}$ (percentage of Purkinje cells, P28)). Note that the percentage of Purkinje cells is relative to control C57BL/6J mice. **b**, Cell type-specific markers (red) and p62 (green) immunofluorescence on sagittal cerebellar sections. Parvalbumin (Parv) was used to identify Purkinje cells and interneurons (stellate and basket cells) in the molecular layer. NeuN was used to distinguish between granule and Golgi cells in the granule cell layer. *En1*^{cre} *Ankrd16*$^{fl/-}$ *Aars*$^{sti/sti}$: Golgi cell (closed arrow head, p62$^+$NeuN$^-$); granule cell (arrow, p62$^+$/NeuN$^+$); basket or stellate cell (open arrow head, p62$^+$/Parv$^+$). $n = 3$ biological replicates. **c**, Ubiquitin (red) and p62 (green) immunofluorescence on sagittal sections of the cortex (layer IV). $n = 3$ biological replicates. Scale bars, 10 μm (**a**), 50 μm (**b** and **c**, low magnification) and 10 μm (**b** and **c**, higher magnification).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Model for the role of ANKRD16 in translational fidelity. a, A point mutation in the editing domain of AlaRS(A734E) results in editing defects, as indicated by deficits in tRNA-independent ATPase activity, which in turn leads to increased levels of incorrectly aminoacylated Ser-tRNA^{Ala}, misincorporation of serine during translation, protein aggregation and cell death. **b**, Interaction of ANKRD16 with the aminoacylation domain of editing-deficient AlaRS(A734E) stimulates tRNA-independent editing, and misactivated serines are transferred onto ANKRD16. Mitigation of serine misactivation prevents *sti*-mediated mistranslation, and thereby prevents protein aggregation and cell death.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0135-x

SWI2/SNF2 ATPase CHR2 remodels pri-miRNAs via Serrate to impede miRNA production

Zhiye Wang[1,2], Zeyang Ma[1,2,5], Claudia Castillo-González[1,2,5], Di Sun[1,2], Yanjun Li[1,2,3], Bin Yu[4], Baoyu Zhao[1], Pingwei Li[1] & Xiuren Zhang[1,2]*

Chromatin remodelling factors (CHRs) typically function to alter chromatin structure. CHRs also reside in ribonucleoprotein complexes, but little is known about their RNA-related functions. Here we show that CHR2 (also known as BRM), the ATPase subunit of the large switch/sucrose non-fermentable (SWI/SNF) complex, is a partner of the Microprocessor component Serrate (SE). CHR2 promotes the transcription of primary microRNA precursors (pri-miRNAs) while repressing miRNA accumulation in vivo. Direct interaction with SE is required for post-transcriptional inhibition of miRNA accumulation by CHR2 but not for its transcriptional activity. CHR2 can directly bind to and unwind pri-miRNAs and inhibit their processing, and this inhibition requires the remodelling and helicase activity of CHR2 in vitro and in vivo. Furthermore, the secondary structures of pri-miRNAs differed between wild-type *Arabidopsis thaliana* and *chr2* mutants. We conclude that CHR2 accesses pri-miRNAs through SE and remodels their secondary structures, preventing downstream processing by DCL1 and HYL1. Our study uncovers pri-miRNAs as a substrate of CHR2, and an additional regulatory layer upstream of Microprocessor activity to control miRNA accumulation.

CHR2 is the ATPase subunit of the large SWI/SNF chromatin-remodelling complex. Such complexes are known to remodel chromatin structures by nucleosome sliding, eviction, or histone variant exchange using energy derived from ATP hydrolysis[1]. Animal CHR2 can also associate with nascent mRNA ribonucleoprotein complexes (pre-mRNPs)[2,3]. The yeast chromatin remodeller ISW1 is involved in the surveillance of nuclear mRNP biogenesis[4]. Some SWI/SNF members can also bind to long-noncoding RNAs (lncRNAs) and participate in assembly of lncRNA-dependent nuclear bodies or regulation of chromatin association[5–8]. These studies suggest that CHR2 and other SWI/SNF subunits have additional roles in RNA biology. Notably, these roles do not require the remodelling activity of the relevant ATPases. Whether and how SWI/SNF factors participate in post-transcriptional processing of RNA is unknown.

MicroRNAs, a large family of small non-coding regulatory RNAs[9], are processed from long pri-miRNAs that contain a hairpin-like fold-back by Microprocessor and Dicing complexes[10,11]. The minimal plant Microprocessor–Dicing complex includes Dicer-like 1 (DCL1), and a double-stranded (ds)RNA-binding protein, Hyponastic leaves 1 (HYL1)[12]. The Microprocessor initially cleaves basal flanking segments of pri-miRNAs to generate precursor miRNAs (pre-miRNAs), and subsequently cuts pre-miRNAs to produce miRNA/* duplexes (where the asterisk represents a miRNA complementary strand)[13]. Although miRNAs are derived from pri-miRNAs, transcript levels of pri-miRNAs do not always correlate with miRNA abundance, possibly owing to various post-transcriptional regulations and processing efficiencies of the RNA species[12,14,15].

SE, a zinc-finger protein, has also been considered to be a core component of plant Microprocessor, because *se* mutations cause pri-miRNA accumulation and miRNA loss in vivo[16]. Although some have argued for a direct role for SE in promoting the enzymatic activity and accuracy of DCL1[17,18], processing of pri-miRNAs by DCL1 in vitro does not seem to require SE[12]; rather, SE might act as a scaffold to recruit the processing machinery including DCL1 and HYL1 to proper RNA substrates to produce miRNAs *in vivo*[12,19,20]. Ars2, the mammalian orthologue of SE, also participates in miRNA-dependent silencing, suggesting that the function of SE has been conserved throughout eukaryotes[21,22]. However, the mechanism by which SE contributes to miRNA production has been unclear.

CHR2 has two roles in miRNA accumulation

We identified CHR2 as a bona fide partner of SE (Fig. 1a, b, Extended Data Fig. 1a–d, Supplementary Information). We then investigated whether CHR2 participated in SE-mediated miRNA production. Adult *chr2-1* mutant plants exhibited short stature and downward curled leaves, whereas *se* mutants had small serrated and upward curled leaves (Extended Data Fig. 1e, f). Small RNA (sRNA) blot assays showed that the amounts of the tested miRNAs, except miR390, were greatly increased in leaves and flowers of *chr2-1*$^{-/-}$ compared to wild-type (Col-0) plants or their heterozygote siblings (*chr2-1*$^{+/-}$) (Fig. 1c, Extended Data Fig. 1g–i). Moreover, both morphological and miRNA abnormalities in *chr2-1* plants were fully complemented by a *PCHR2-gCHR2-Flag-4Myc* (*PCHR2-gCHR2-FM*) transgene, showing that the lack of CHR2 causes both morphological and miRNA abnormalities (Extended Data Fig. 1j, k).

We compared miRNA profiles in *chr2-1* and Col-0 plants through deep-sequencing analysis. We normalized sequencing read counts to miR390, which consistently showed no changes between *chr2-1* and Col-0 plants in either sRNA blots or sequencing (see Methods). We found that 114 of 365 annotated miRNAs were upregulated at least 1.5-fold in *chr2-1* relative to Col-0 plants (Fig. 1d, Supplementary Table 1). We focused on the 259 canonical SE-dependent miRNAs and

[1]Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX, USA. [2]Institute for Plant Genomics and Biotechnology, Texas A&M University, College Station, TX, USA. [3]College of Plant Protection, Nanjing Agricultural University, Nanjing, China. [4]School of Biological Sciences & Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE, USA. [5]These authors contributed equally: Zeyang Ma, Claudia Castillo-González. *e-mail: xiuren.zhang@tamu.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

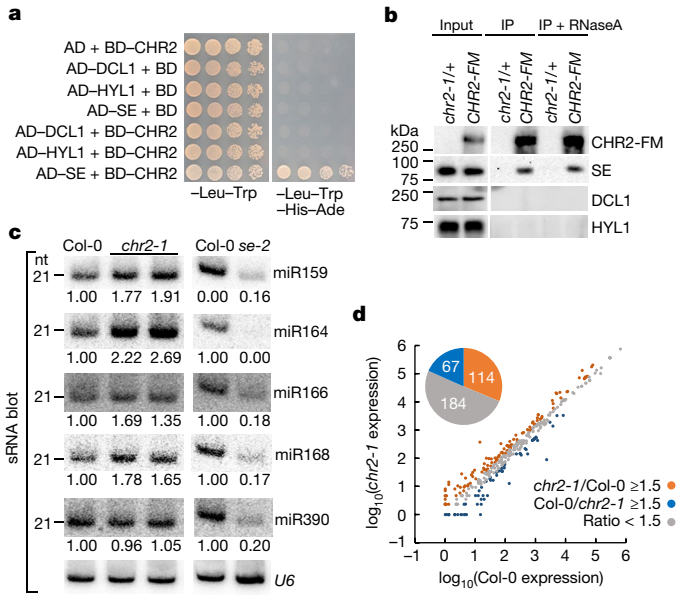


Fig. 1 | CHR2 is a novel partner of SE and represses miRNA accumulation. a, **b**, Y2H (**a**) and co-immunoprecipitation (IP) (**b**) assays show the specific CHR2–SE interaction. AD, GAL4-activation domain; BD, GAL4-DNA-binding domain; AD or BD-protein, AD or BD-fusion proteins; *CHR2-FM*, *chr2-1;PCHR2-gCHR2–Flag–4Myc*. **c**, sRNA blot analysis shows that miRNAs accumulated in *chr2-1* mutant plants. **d**, sRNA sequencing analysis shows that a substantial number of miRNAs accumulated in *chr2-1* mutants. Compared to Col-0, miRNAs with at least 1.5-fold higher (*chr2-1*/Col-0 ≥ 1.5) or lower (Col-0/*chr2-1* ≥ 1.5) expression in *chr2-1* are indicated by orange dots and blue dots, respectively. The grey dots indicate differences in expression level < 1.5-fold (ratio < 1.5).

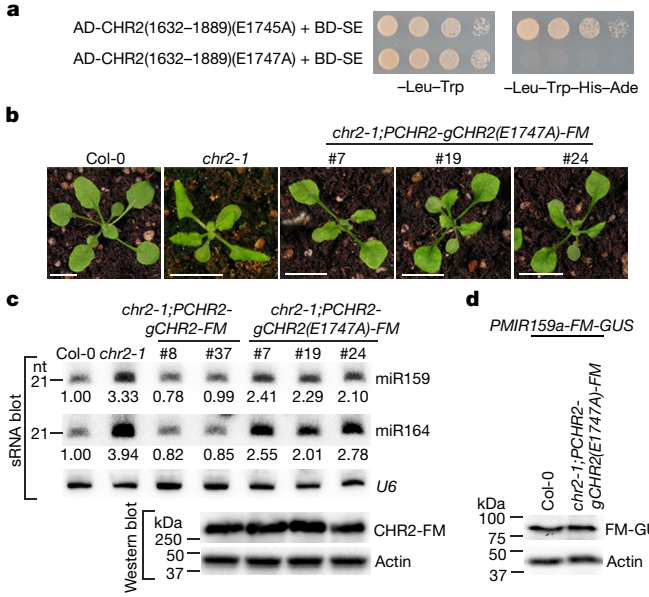


Fig. 2 | Uncoupling of transcriptional and post-transcriptional regulatory roles of CHR2 in miRNA biogenesis. a, Y2H assay shows that CHR2(E1747) is required for its interaction with SE. **b**, **c**, CHR2(E1747A) partially rescued the morphological and miRNA accumulation defects seen in *chr2-1* plants. **b**, Scale bars, 1 cm. **c**, sRNA blot analysis (top) and western blot analysis of CHR2 or its variant (bottom). **d**, Western blot analysis of GUS protein from *PMIR159a-FM-GUS* homozygotes show that CHR2(E1747A) does not alter the transcriptional level of *MIR159a* in Col-0 and *chr2-1;PCHR2-gCHR2(E1747A)-FM* backgrounds ($n > 30$).

found that approximately one-quarter of these SE-dependent miRNAs, representing about 59% of CHR2-repressed miRNAs, were significantly enhanced in *chr2-1* plants (Extended Data Fig. 1l, m, Supplementary Table 1). Thus, CHR2 is a previously unrecognized negative regulator of miRNA accumulation for a large subset of miRNAs in *Arabidopsis*.

We then investigated the transcription levels of miRNA genes (MIRs) in *chr2* mutants through *PMIR-FM-GUS* reporter assays and measurement of nascent pri-miRNA levels. Notably, unlike the effects on miRNA accumulation, MIR gene transcription was reduced in *chr2* mutants (Extended Data Fig. 2, Supplementary Table 2, Supplementary Information). Thus, CHR2 must have an additional inhibitory role in miRNA biogenesis, beyond its canonical function in promoting MIR transcription.

Uncoupled CHR2 roles in miRNA biogenesis

To study whether the functions of CHR2 in miRNA biogenesis depend on SE, we mapped the CHR2–SE interaction interface. Using yeast two-hybrid (Y2H) assays with dozens of CHR2 mutants, we found that CHR2(E1747A) abolished the CHR2–SE interaction (Fig. 2a, Extended Data Fig. 3a–e, Supplementary Information). Thus, E1747 of CHR2 is essential for the specific CHR2–SE interaction. We then transformed *PCHR2-gCHR2(E1747A)-FM* into *chr2-1* plants and obtained T2 homozygotes. Notably, CHR2(E1747A) only partially rescued the morphological and miRNA defects in *chr2-1* plants (Fig. 2b, c). This result indicated that the direct CHR2–SE interaction is required for inhibition of miRNA accumulation in vivo by CHR2.

Through genetic crossing, we obtained *chr2-1;PCHR2-gCHR2(E1747A)-FM;PMIR159a-FM-GUS* homozygotes. The expression level of the *GUS* reporter was comparable in *chr2-1;PCHR2-gCHR2(E1747A)-FM* and Col-0 backgrounds (Fig. 2d, Extended Data Fig. 3f). Thus, the E1747A mutation does not alter the transcriptional function of CHR2. This result also suggests that enhanced miRNA accumulation in the *chr2-1;PCHR2-gCHR2(E1747A)-FM* hypomorphic line results from compromised inhibition of SE-mediated miRNA biogenesis by CHR2(E1747A). Thus, the transcriptional and post-transcriptional regulatory roles of CHR2 in miRNA biogenesis can be uncoupled via SE.

CHR2 inhibits pri-miRNA processing in vitro

CHR2 does not affect the expression of miRNA metabolism components or SE-dependent pre-mRNA splicing (Extended Data Fig. 4, Supplementary Table 3, Supplementary Information). To test whether CHR2 directly inhibits SE-mediated pri-miRNA processing, we established an in vitro Microprocessor system using recombinant DCL1, SE, and HYL1. The recombinant Microprocessor (DCL1–HYL1–SE) cleaved pri-miR166f as efficiently and accurately as DCL1 immunoprecipitates from plants[12] (Extended Data Fig. 5a, b). Thus, the recombinant Microprocessor can recapitulate in vivo miRNA biogenesis. Notably, incubation of purified CHR2, but not a control (hSTING)[23], with pri-miRNAs before application of HYL1–SE substantially inhibited Microprocessor activity (Fig. 3a, Extended Data Fig. 5c). CHR2-mediated inhibition of Microprocessor activity was concentration-dependent and also enhanced by prolonged incubation with pri-miRNA (Extended Data Fig. 5d, e). By contrast, treatment with CHR2 after pri-miRNA incubation with HYL1–SE only slightly inhibited DCL1 activity (Fig. 3b, Extended Data Fig. 5f). Thus, Microprocessor, when loaded with pri-miRNAs, could largely bypass CHR2 inhibition in vitro. This result also suggests that CHR2 impedes miRNA biogenesis by acting upstream of Microprocessor activity in vivo.

Microprocessor cleavage activity was also substantially inhibited by CHR2(E1747A) that was pre-incubated with pri-miRNAs, but to a lesser extent than by CHR2 (Fig. 3c, Extended Data Fig. 5g). This result indicates that CHR2-mediated inhibition of Microprocessor activity in vitro results from features other than its SE-binding ability. By contrast, addition of CHR2(E1747A) to pri-miRNAs that had been pre-incubated with SE had only a marginal effect on miRNA production (Fig. 3d, Extended Data Fig. 5h). Thus, if SE binds to pri-miRNA before CHR2, obstruction of the CHR2–SE interaction compromises CHR2-mediated inhibition of pri-miRNA processing in vitro. Notably,

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

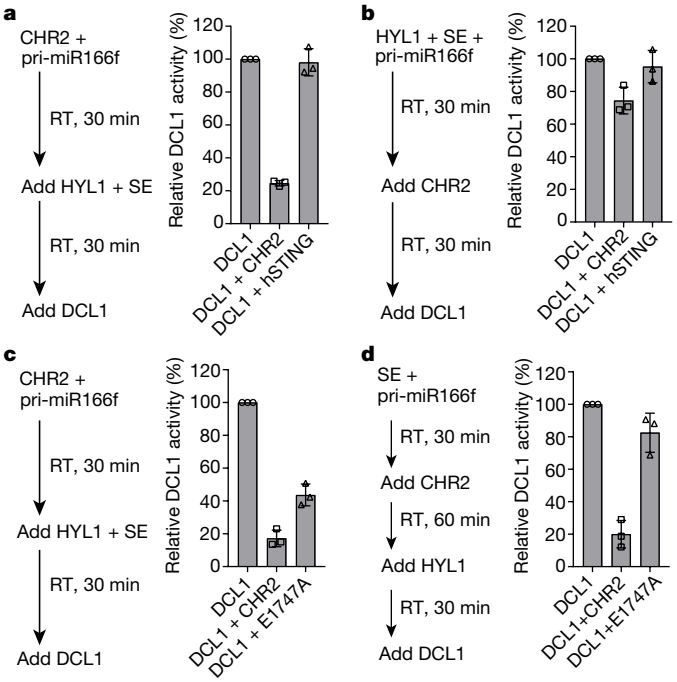


Fig. 3 | In vitro Microprocessor assays show that CHR2 inhibits pri-miRNA processing. a, Pre-incubation of CHR2 with pri-miRNAs inhibited pri-miRNA processing. **b**, Pre-incubation of HYL1–SE with pri-miRNAs largely bypassed the CHR2-mediated inhibition of pri-miRNA processing. **c**, Pre-incubation of pri-miRNA with CHR2(E1747A) substantially reduced pri-miRNA processing. **d**, CHR2(E1747A) marginally inhibited pri-miRNA processing when pri-miRNA was pre-incubated with SE. Molar ratio of SE:CHR2 (or CHR2(E1747A)) was 4:1. **a**–**d**, Left, reconstitution orders; right, quantification of relative cleavage efficiency with s.d. from three replicates.

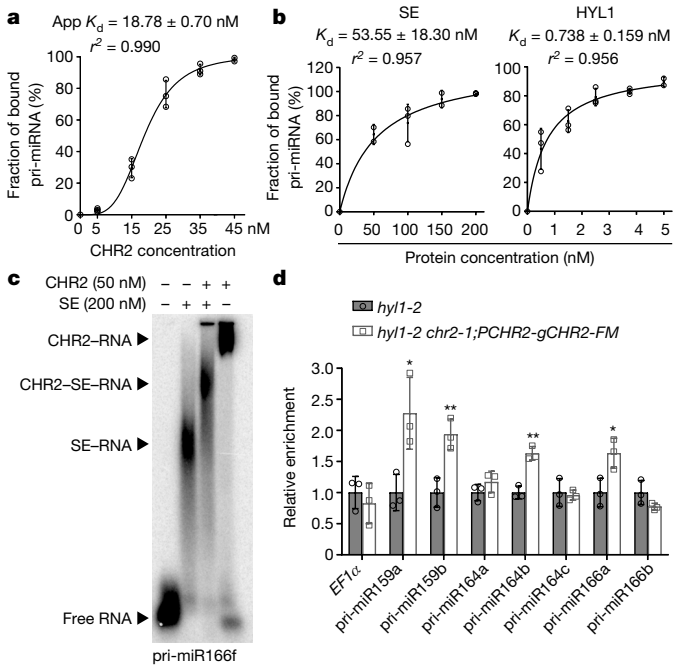


Fig. 4 | CHR2 binds to pri-miRNAs in vitro and in vivo. a, **b**, The binding curves of CHR2 (**a**), SE and HYL1 (**b**) to pri-miRNA. The (apparent) K_d values were calculated from the EMSA image quantification with s.d. from three experiments. **c**, EMSA shows that CHR2, SE and pri-miRNA form a ribonucleoprotein complex. **d**, RIP assay shows that CHR2 binds to pri-miRNAs in vivo. The relative signal of pri-miRNAs was calculated with s.d. from three biological repeats (*$P < 0.05$; **$P < 0.01$; unpaired, two-tailed Student's *t*-test).

these results are consistent with the observation that CHR2 requires SE to inhibit miRNA accumulation in vivo (Fig. 2c). Thus, we conclude that CHR2 obtains pri-miRNAs from SE to fulfil its inhibition of pri-miRNA processing.

CHR2 binds to pri-miRNAs

CHR2 does not block access of DCL1–HYL1 to SE to process pri-miRNAs (Model 1, Extended Data Figs. 6, 7, Supplementary Information). Next, we asked whether CHR2 sequestered pri-miRNAs from SE, thwarting their handover to DCL1–HYL1 (Model 2, Extended Data Fig. 6a). Electrophoretic mobility shift assays (EMSAs) showed that CHR2 bound to pri-miRNAs and pre-miRNAs (apparent dissociation constant (K_d) = 18.78 ± 0.70 nM; Hill coefficient (n_H) = 3.87) as strongly as to double-stranded DNA (dsDNA) (apparent $K_d \approx 16$ nM, $n_H = 4.27$) (Fig. 4a, Extended Data Fig. 8a–f). The sigmoidal CHR2–nucleic acid binding curve (Fig. 4a, Extended Data Fig. 8f) suggests cooperativity between multiple nucleic acid binding sites in CHR2[24] in substrate binding.

SE and HYL1 can also bind to pri-miRNAs[18–20] with binding affinities of approximately 50 and 0.7 nM, respectively (Fig. 4b, Extended Data Fig. 8g, h). These results indicate that CHR2 binds to pri-miRNAs with a higher affinity than SE but with a substantially lower affinity than HYL1. The results also suggest that CHR2 has the potential to compete with SE, but not with HYL1, for pri-miRNAs. However, EMSA showed that co-incubation of CHR2 and SE with pri-miRNAs resulted in a shifting pattern different from the mobility of either CHR2–pri-miRNA or SE–pri-miRNA complexes (Fig. 4c). These results suggest that CHR2, SE, and pri-miRNA can form stable complexes with a distinct electrophoretic mobility. Alternatively, CHR2 may alter the pri-miRNA structure and thus, the mobility of the SE–pri-miRNA complex. By contrast, whenever HYL1 was applied, individually or combined with other proteins, the mobility of the pri-miRNA–protein complexes was always identical to that of HYL1–pri-miRNA alone (Extended Data Fig. 8i). This result suggests that HYL1 could readily sequester pri-miRNAs from SE and/or CHR2 in vitro, and that CHR2 and SE could not preclude pri-miRNA handover to DCL1–HYL1 for processing. This result is also consistent with the observation that inhibition of Microprocessor activity by CHR2 can be largely attenuated when pri-miRNAs are pre-incubated with HYL1–SE (Fig. 3b).

Next, we performed ribonucleoprotein immunoprecipitation (RIP) experiments using *hyl1-2 chr2-1;PCHR2-gCHR2-FM* homozygotes containing unprocessed pri-miRNAs (Extended Data Fig. 8j). The RIP results showed that pri-miR159a, pri-miR159b, pri-miR164b, and pri-miR166a, which are the main contributors to their respective mature miRNAs, were all significantly enriched in the CHR2 immunoprecipitate, whereas this pattern was not observed for other pri-miRNAs (pri-miR164a and pri-miR164c) that do not contribute substantially to mature miRNAs (Fig. 4d). Moreover, treatment of the samples with RNases before reverse transcription did not yield any signals, suggesting that the enrichment was specifically from CHR2-bound RNAs (Extended Data Fig. 8k). Together, these results validated the theory that CHR2 binds to many species of nascent pri-miRNAs in vivo, further suggesting that CHR2 acts on SE-bound pri-miRNAs before their handover to DCL1–HYL1 in vivo.

CHR2 remodels SE-bound pri-miRNAs

Plants containing the hypomorphic *chr2-1a* allele with compromised ATPase activity retained enhanced miRNA accumulation compared to Col-0 plants (Extended Data Fig. 2d–g). Thus, inhibition of miRNA accumulation by CHR2 involves its ATPase activity. We next hypothesized that CHR2 might unwind self-complementary pri-miRNAs and alter their hairpin structures (Model 3, Extended Data Fig. 6a). To test this hypothesis, we conducted CHR2 remodelling assays with ^{32}P-labelled but nicked pri-miR166f. CHR2, but not HYL1, clearly released a large portion of the 26-nucleotide (nt) RNA fragment from the nicked pri-miR166f (Fig. 5a, Extended Data Fig. 9a–d). Therefore,

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

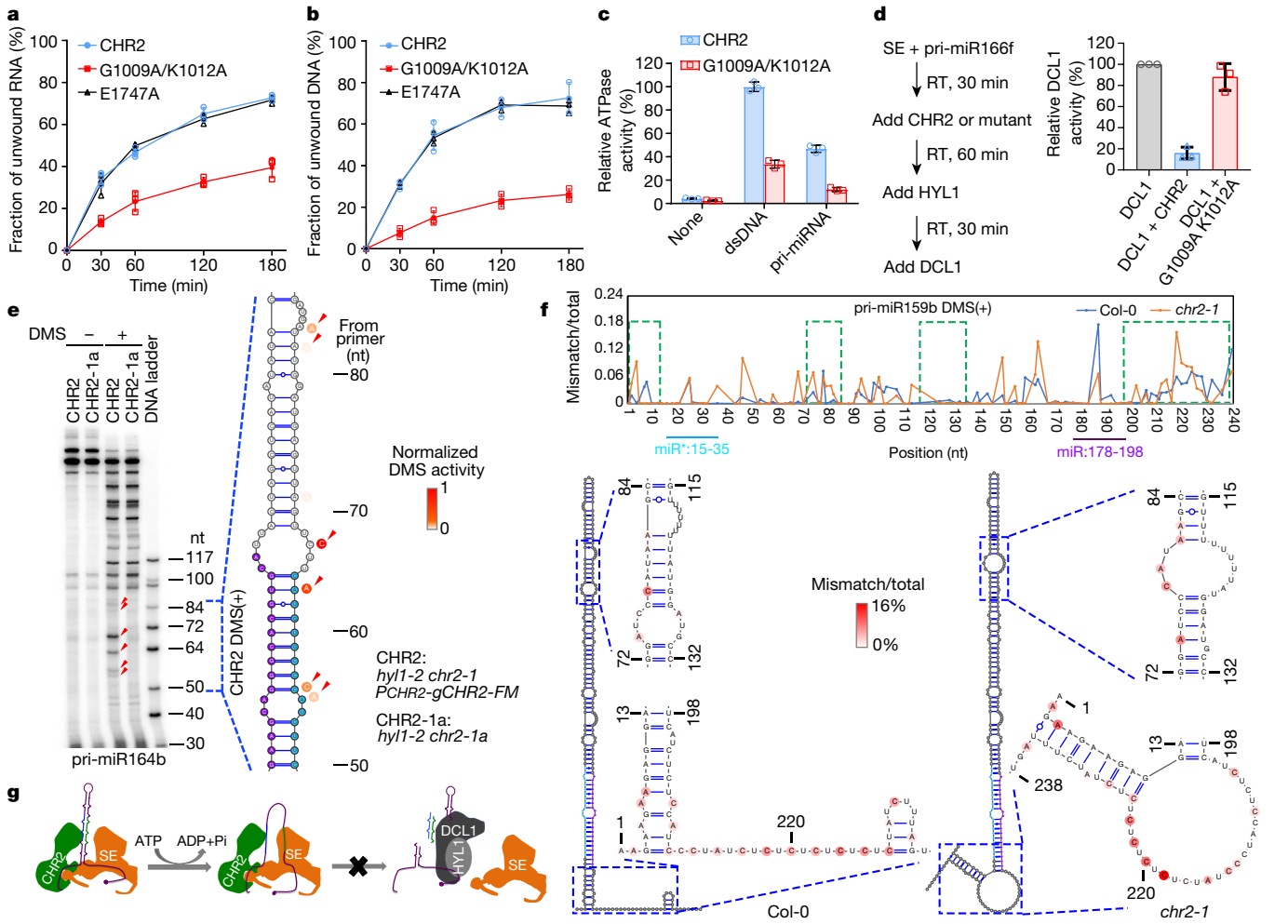


Fig. 5 | CHR2 remodels pri-miRNAs to prevent their processing. **a**, **b**, Time-course of remodelling and helicase activities of CHR2 and variants on pri-miRNAs (**a**) and dsDNA (**b**). **c**, Thin layer chromatography (TLC) assays shows ATPase activities of CHR2 and CHR2(G1009A/K1012A) with or without dsDNA or pri-miRNA. **d**, Right, Microprocessor assay shows that CHR2(G1009A/K1012A) did not inhibit pri-miRNA processing (right). Left, reconstitution order. In **a**–**d**, image quantifications were calculated with s.d. from three replicates. **e**, DMS-primer extension assays show nucleotides (red arrows) in pri-miR164b with stronger DMS activities in *hyl1-2 chr2-1;PCHR2-gCHR2-FM* plants than in *hyl1-2 chr2-1a* plants. **f**, DMS–MaPseq analysis shows that CHR2 alters nucleotide pairings in upper stem and base region of pri-miR159b. Top, average mutation frequencies of A and C are plotted along pri-miRNA sequence. The regions within green dashed boxes correspond to zoomed-in secondary structures modelled from the DMS activities (bottom). Colour-coded nucleotides had different DMS activities between Col-0 and *chr2-1* plants. miRNA and * strands are shown in purple and cyan, respectively. **g**, Proposed model for inhibition of pri-miRNA processing by CHR2.

CHR2 does unwind pri-miRNAs, in a way similar to dsDNA (Fig. 5b, Extended Data Fig. 9e, f). These results were unexpected, as SWI/SNF complexes are mainly thought to lack helicase activity[1,25].

Similar to other SWI2/SNF2 ATPases[26,27], CHR2 had only basal ATPase activity in the ground state. Nonetheless, dsDNA and pri-miRNA stimulated its ATPase activity by approximately 25-fold and 12-fold, respectively (Fig. 5c, Extended Data Fig. 9g–i). The lower stimulation of ATPase activity by pri-miRNA relative to dsDNA was probably because it harbours mispairings and internal loops. Notably, mutations in the CHR2 ATP-binding site (G1009A/K1012A or N1392A/R1417A) diminished the dsDNA- and/or pri-miRNA-stimulating ATPase activities of CHR2 by threefold (Fig. 5c, Extended Data Fig. 2e, 9g–i). Consistently, CHR2(G1009A/K1012A) and the CHR2 nucleotide-binding mutant (CHR2(D1355A/R1385A)), but not CHR2(E1747A), also strongly reduced the CHR2 remodelling/helicase activity (Fig. 5a, b, Extended Data Fig. 9c–f). These results indicate that the unwinding of dsDNA and pri-miRNAs by CHR2 is driven by ATP hydrolysis. In addition, the CHR2 mutants with alterations in the ATP- and nucleotide-binding sites, but not the E1747A mutant, had slightly reduced affinity for pri-miRNAs relative to wild-type CHR2 (Extended Data Fig. 10a). Thus, the pri-miRNA binding and remodelling activities of CHR2 depend on its ATPase domain, but not on the SE interaction interface.

We further performed Microprocessor assays using the remodelling-defective CHR2(G1009A/K1012A). As CHR2 is likely to access pri-miRNAs through SE in vivo (Fig. 2c), we pre-incubated SE protein with pri-miRNAs before the application of CHR2 or CHR2(G1009A/K1012A). Whereas CHR2 substantially inhibited Microprocessor activity, CHR2(G1009A/K1012A) had negligible effects on pri-miRNA processing (Fig. 5d, Extended Data Fig. 10b). The loss of inhibition of pri-miRNA processing was probably not caused by the slightly reduced affinity of CHR2(G1009A/K1012A) for pri-miRNAs compared to wild-type CHR2 (apparent $K_d \approx 19$ nM and 28 nM, respectively) because CHR2 and SE are in the same complex with pri-miRNAs (Fig. 4c). Rather, the result indicates that the CHR2 remodelling activity is critical for inhibition of pri-miRNA processing in vitro.

To study how the remodelling mutations affect miRNA accumulation in vivo, we created *chr2-1;PCHR2-gCHR2(G1009A/K1012A)-FM* and *chr2-1;PCHR2-gCHR2(D1355A/R1385A)-FM* transgenic lines. Notably, the morphological and miRNA accumulation defects seen in *chr2-1* plants were partially rescued in these transgenic plants (Extended Data Fig. 10c–e). We also introduced the *PMIR159b-FM-GUS* transgene into

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

the *chr2-1*;*PCHR2-gCHR2(G1009A/K1012A)-FM* hypomorphic line by crossing. Notably, the CHR2 variant reduced MIR gene transcription in *chr2-1*;*PCHR2-gCHR2(G1009A/K1012A)-FM*;*PMIR159b-FM-GUS* homozygotes (Extended Data Fig. 10f). Thus, the increase in miRNA abundance in the partial complementation lines should result from compromised post-transcriptional inhibition of pri-miRNA processing by CHR2 variants that lack the remodelling function.

CHR2 alters pri–miRNA folding in vivo

To determine whether CHR2 remodels pri-miRNAs in vivo, we probed RNA secondary structures through a dimethyl sulfide (DMS)-based method[28] (Extended Data Fig. 11a–c). DMS methylates the base-pairing faces of adenosine (A) and cytidine (C) of RNA in loops, bulges, mismatches and joining regions, and such methylation precludes the reverse transcription reaction. Strikingly, the reactivity of DMS with pri-miR164b differed between transgenic plants expressing wild-type *CHR2* and those expressing *CHR2-1a*, which has defective ATPase activity (Fig. 5e, Extended Data Fig. 2d–g, 11d). Enhanced DMS modifications, detected by premature termination of reverse transcription, were observed in the miRNA/* duplex and the proximal upper stem of pri-miR164b in *CHR2* plants compared to *chr2-1a* plants. This result indicates that there are more unpaired or unprotected nucleotides in the duplex region of pri-miR164b in a *CHR2* background than in the *chr2* mutant.

Finally, we adopted a method of DMS mutational profiling with Illumina sequencing (DMS–MaPseq) to target multiple RNAs in vivo[29] (Extended Data Fig. 11a). DMS–MaPseq uses a special reverse transcriptase enzyme from a group II self-splicing intron (TGIRT)[30]. TGIRT can read through the DMS-modified A and C template bases and insert mismatches into the corresponding cDNAs, thereby providing signatures for mapping of DMS modifications along reverse transcription products[29]. We targeted 16 pri-miRNAs, and eight were successfully and consistently amplified between two biological replicates. Notably, DMS–MaPseq showed meaningful differences in DMS reactivity of the eight pri-miRNAs, but not the control *UBQ4* mRNA, between Col-0 and *chr2-1* plants (Fig. 5f, Extended Data Fig. 11e–h, 12). First, pri-miR166a and pri-miR168a tended to form shortened upper stems and branched terminal loops in Col-0 plants, whereas they displayed extended upper stems and linear terminal loops in *chr2-1* plants (Extended Data Fig. 11g, h). As multi-branched terminal loops and shortened upper stems can cause loop-to-base abortive processing of pri-miRNAs whereas elongated upper stems promote productive processing of pri-miRNAs[12], CHR2 might remodel the structures of terminal loops and upper stems to control the productive processing of pri-miRNAs. Second, modelling of pri-miR168b folding showed that the lower stem was increased from 9 nt in Col-0 plants to 16 nt in *chr2-1* plants (Extended Data Fig. 12a). Such an increase places miR168b/* an ideal distance of 15–17 nt from a reference single-strand basal loop region[12], allowing accurate processing of pri-miR168b. Finally, pri-miR159 and pri-miR319 are sequentially processed from terminal loop to base to eventually release miRNA/*[31]. In *chr2-1* plants, pri-miR159b and pri-miR319b had stronger DMS reactivity in the upper stems, suggesting lesser folding or more unpaired nucleotides in these regions and potentially increasing pri-miRNA processing from the terminal loop to the lower base (Fig. 5f; Extended Data Fig. 12b). Furthermore, the lower stem of pri-miR159b was also altered so that miR159/* was immediately adjacent to a bigger and more basal loop in *chr2-1* plants (Fig. 5f). This conformation is likely to further promote accurate and efficient processing of miRNA/* from pre-miRNAs, as both animal Dicer and plant DCL1 recognize the loop–bulge structures in addition to the 5′ and 3′ ends of pre-miRNAs for accurate processing[12,32]. Thus, CHR2 alters the in vivo secondary structures of pri-miRNAs or, possibly, their association with cellular components in various ways.

Discussion

Here we report a non-canonical function of CHR2 in miRNA biogenesis: while CHR2 positively regulates transcription of MIR loci owing to its conventional chromatin remodelling activity, it can also directly inhibit post-transcriptional processing of pri-miRNAs. On the basis of our results, we propose a model in which CHR2 inhibits miRNA production by obtaining pri-miRNAs from SE and remodelling the RNA substrates, rendering them unsuitable for processing by Microprocessor (Fig. 5g). Thus, the SWI2/SNF2 ATPase goes beyond its canonical substrates of nucleosome[1] and Polycomb complexes[33], and acts directly on its new substrate of pri-miRNAs. The biological significance of these self-opposing functions in miRNA biogenesis is that CHR2 provides a new and balanced regulatory mechanism to stabilize the supply of proper substrates upstream of Microprocessor activity. In this scenario, CHR2–SE may serve as a specific reader and modifier of sequences and structures of pri-miRNAs and fine-tune miRNA biogenesis *in vivo*. This mechanism may also partially explain frequently observed inconsistencies between the expression of MIR genes and the abundance of mature miRNAs[14,15].

Several cellular components, including transcription factors, can increase or decrease Microprocessor activity by modulating or sequestering core or regulatory Microprocessor components[34–37]. However, CHR2 can act as both a positive and a negative regulator of miRNA biogenesis, and the net outcome depends on the balance of two opposite forces. Thus, CHR2 represents a new paradigm for the regulation of miRNA biogenesis by exerting opposing effects in consecutive biological processes (pri-miRNA transcription and processing) through the conformation of different protein complexes (SWI/SNF and CHR2–SE), acting on different molecular substrates (chromatin and structured RNA, respectively). As CHR2 and SE/Ars2 are conserved throughout eukaryotes, it would be interesting to learn whether such a mechanism exists beyond *Arabidopsis*. Moreover, whether CHR2 and SE/Ars2 participate in remodelling of pre-mRNPs and lncRNA complexes besides pri-miRNAs would also be an exciting topic for future investigation.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0135-x.

Received: 21 September 2017; Accepted: 9 March 2018;
Published online: 16 May 2018

1. Clapier, C. R., Iwasa, J., Cairns, B. R. & Peterson, C. L. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nat. Rev. Mol. Cell Biol.* **18**, 407–422 (2017).
2. Batsché, E., Yaniv, M. & Muchardt, C. The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat. Struct. Mol. Biol.* **13**, 22–29 (2006).
3. Tyagi, A., Ryme, J., Brodin, D., Ostlund Farrants, A. K. & Visa, N. SWI/SNF associates with nascent pre-mRNPs and regulates alternative pre-mRNA processing. *PLoS Genet.* **5**, e1000470 (2009).
4. Babour, A. et al. The chromatin remodeler ISW1 is a quality control factor that surveys nuclear mRNP biogenesis. *Cell* **167**, 1201–1214.e15 (2016).
5. Kawaguchi, T. et al. SWI/SNF chromatin-remodeling complexes function in noncoding RNA-dependent assembly of nuclear bodies. *Proc. Natl Acad. Sci. USA* **112**, 4304–4309 (2015).
6. Han, P. et al. A long noncoding RNA protects the heart from pathological hypertrophy. *Nature* **514**, 102–106 (2014).
7. Zhu, P. et al. LncBRM initiates YAP1 signalling activation to drive self-renewal of liver cancer stem cells. *Nat. Commun.* **7**, 13608 (2016).
8. Prensner, J. R. et al. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genet.* **45**, 1392–1398 (2013).
9. Li, S., Castillo-González, C., Yu, B. & Zhang, X. The functions of plant small RNAs in development and in stress responses. *Plant J.* **90**, 654–670 (2017).
10. Nguyen, T. A. et al. Functional anatomy of the human Microprocessor. *Cell* **161**, 1374–1387 (2015).
11. Ha, M. & Kim, V. N. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15**, 509–524 (2014).
12. Zhu, H. et al. Bidirectional processing of pri-miRNAs with branched terminal loops by *Arabidopsis* Dicer-like1. *Nat. Struct. Mol. Biol.* **20**, 1106–1115 (2013).
13. Rogers, K. & Chen, X. Biogenesis, turnover, and mode of action of plant microRNAs. *Plant Cell* **25**, 2383–2399 (2013).
14. Barciszewska-Pacak, M. et al. *Arabidopsis* microRNA expression regulation in a wide range of abiotic stress responses. *Front. Plant Sci.* **6**, 410 (2015).
15. Thomson, J. M. et al. Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes Dev.* **20**, 2202–2207 (2006).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

16. Grigg, S. P., Canales, C., Hay, A. & Tsiantis, M. SERRATE coordinates shoot meristem function and leaf axial patterning in *Arabidopsis*. *Nature* **437**, 1022–1026 (2005).
17. Dong, Z., Han, M. H. & Fedoroff, N. The RNA-binding proteins HYL1 and SE promote accurate *in vitro* processing of pri-miRNA by DCL1. *Proc. Natl Acad. Sci. USA* **105**, 9970–9975 (2008).
18. Iwata, Y., Takahashi, M., Fedoroff, N. V. & Hamdan, S. M. Dissecting the interactions of SERRATE with RNA and DICER-LIKE 1 in *Arabidopsis* microRNA precursor processing. *Nucleic Acids Res.* **41**, 9129–9140 (2013).
19. Machida, S., Chen, H. Y. & Adam Yuan, Y. Molecular insights into miRNA processing by *Arabidopsis* thaliana SERRATE. *Nucleic Acids Res.* **39**, 7828–7836 (2011).
20. Yang, S. W. et al. Structure of *Arabidopsis* HYPONASTIC LEAVES1 and its molecular implications for miRNA processing. *Structure* **18**, 594–605 (2010).
21. Gruber, J. J. et al. Ars2 links the nuclear cap-binding complex to RNA interference and cell proliferation. *Cell* **138**, 328–339 (2009).
22. Sabin, L. R. et al. Ars2 regulates both miRNA- and siRNA- dependent silencing and suppresses RNA virus infection in *Drosophila*. *Cell* **138**, 340–351 (2009).
23. Shu, C., Yi, G., Watts, T., Kao, C. C. & Li, P. Structure of STING bound to cyclic di-GMP reveals the mechanism of cyclic dinucleotide recognition by the immune system. *Nat. Struct. Mol. Biol.* **19**, 722–724 (2012).
24. Farrona, S., Hurtado, L. & Reyes, J. C. A nucleosome interaction module is required for normal function of *Arabidopsis thaliana* BRAHMA. *J. Mol. Biol.* **373**, 240–250 (2007).
25. Côté, J., Quinn, J., Workman, J. L. & Peterson, C. L. Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. *Science* **265**, 53–60 (1994).
26. Liu, X., Li, M., Xia, X., Li, X. & Chen, Z. Mechanism of chromatin remodelling revealed by the Snf2-nucleosome structure. *Nature* **544**, 440–445 (2017).
27. Dürr, H., Körner, C., Müller, M., Hickmann, V. & Hopfner, K. P. X-ray structures of the *Sulfolobus solfataricus* SWI2/SNF2 ATPase core and its complex with DNA. *Cell* **121**, 363–373 (2005).
28. Ding, Y. et al. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).
29. Zubradt, M. et al. DMS-MaPseq for genome-wide or targeted RNA structure probing *in vivo*. *Nat. Methods* **14**, 75–82 (2017).
30. Mohr, S. et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19**, 958–970 (2013).
31. Bologna, N. G., Mateos, J. L., Bresso, E. G. & Palatnik, J. F. A loop-to-base processing mechanism underlies the biogenesis of plant microRNAs miR319 and miR159. *EMBO J.* **28**, 3646–3656 (2009).
32. Gu, S. et al. The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing *in vivo*. *Cell* **151**, 900–911 (2012).
33. Kadoch, C. et al. Dynamics of BAF-Polycomb complex opposition on heterochromatin in normal and oncogenic states. *Nat. Genet.* **49**, 213–222 (2017).
34. Davis, B. N., Hilyard, A. C., Lagna, G. & Hata, A. SMAD proteins control DROSHA-mediated microRNA maturation. *Nature* **454**, 56–61 (2008).
35. Suzuki, H. I. et al. Modulation of microRNA processing by p53. *Nature* **460**, 529–533 (2009).
36. Cheng, T. L. et al. MeCP2 suppresses nuclear microRNA processing and dendritic growth by regulating the DGCR8/Drosha complex. *Dev. Cell* **28**, 547–560 (2014).
37. Mori, M. et al. Hippo signaling regulates microprocessor and links cell-density-dependent miRNA biogenesis to cancer. *Cell* **156**, 893–906 (2014).

Acknowledgements We thank A. Jerzmanowski and Y. Cui for sharing *CHR2* plasmids and *chr2-1* alleles; A. Millar for *PMIR159a-GUS* and *PMIR159b-GUS* lines; P. He and P. Manavella for Y3H constructs; C. Kaplan and T. Devarenne for reading the manuscript; R. Hays for technical advice; and L. Zeng and F. Raushel for equipment and facility support. This work was supported by the NSF grant MCB-1716243 and CPRIT grant RP160822 to X.Z. D.S. and Y.L. were supported by the China Scholar Council Fellowship.

Reviewer information *Nature* thanks K. P. Hopfner, Z. Szweykowska-Kulińska and P. Waterhouse for their contribution to the peer review of this work.

Author contributions X.Z. conceptualized the study. X.Z., Z.W. and C.C.-G. designed experiments. X.Z. performed gel filtration chromatography and immunoprecipitation with mass spectrometry. Z.W. performed all biochemistry assays, genetic and molecular studies, RIP, in vivo pri-miRNA structure probing, and part of the analysis of high-throughput sequencing data. Z.M. performed subcellular localization assays and analysed high-throughput sequencing data. C.C.-G. performed protein alignment and provided intellectual input. D.S. performed LCI assay. Y.L. performed part of the Y2H and Y3H and genotyping assays. B.Y. provided antibodies and additional reagents. B.Z. and P.L. provided the baculovirus/insect expression system and intellectual input. X.Z. and Z.W. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0135-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0135-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to X.Z.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Plant materials and growth conditions. *Arabidopsis thaliana* ecotype Columbia (Col-0), *chr2-1* (SALK_030046), *se-1* (CS3257), *se-2* (SAIL_44_G12), *se-3* (SALK_083196), *hyl1-2* (SALK_064863) and *dcl1-9* (CS3828) were used for this study.

For generating complementation and point mutation (partial complementation) lines, the heterozygous *chr2-1*$^{+/-}$ mutant was transformed with binary vectors of PCHR2-gCHR2-FM or its derivatives by the floral dip transformation method[38]. In T2 transgenic plants, the homozygous *chr2-1* background was genotyped by PCR using specific primers (Supplementary Table 4); transgenic plants harbouring dual-tagged CHR2 and its derivatives were screened by western blot analysis. The pleiotropic phenotype of *chr2-1* was fully rescued in the T2 progeny of the *PCHR2-gCHR2-FM* transgenic plants, indicating that dual-tagged CHR2 is as functionally active as wild-type CHR2.

Double mutants such as *chr2-1a se-3*, *chr2-1a hyl1-2* and *chr2-1a dcl1-9* were obtained by crossing the partial complementation line of *chr2-1;PCHR2-gCHR2(N1392A/R1417A)-FM* #49 (termed *chr2-1a*) with heterozygous *se-3*, homozygous *hyl1-2* and heterozygous *dcl1-9* lines respectively. In the F2 generation, homozygous double mutants were identified by PCR using primers listed in Supplementary Table 4.

The *hyl1-2 chr2-1;PCHR2-gCHR2-FM* material was obtained by crossing the complementation line of *chr2-1;PCHR2-gCHR2-FM* #8 with *hyl1-2* plants. In the F2 generation, homozygous *hyl1-2 chr2-1;PCHR2-gCHR2-FM* plants were identified by PCR using primers listed in Supplementary Table 4.

For GUS histochemical analysis, heterozygous *chr2-1*$^{+/-}$ plants were transformed with the MIR promoter-driven *FM-GUS* (PMIR159a-FM-GUS, PMIR159b-FM-GUS, PMIR164a-FM-GUS, PMIR164b-FM-GUS and PMIR164c-FM-GUS vectors) by floral transformation[38]. In T2 transgenic plants, homozygous *chr2-1*$^{-/-}$ plants were identified by phenotypic segregation. *chr2-1;PCHR2-gCHR2(E1747A)-FM;PMIR159a-FM-GUS* and *chr2-1;PCHR2-gCHR2(G1009A/K1012A)-FM;PMIR159b-FM-GUS* plants were obtained by crossing *chr2-1*$^{+/-}$*;PCHR2-gCHR2(E1747A)-FM* #7 and *chr2-1*; *PCHR2-gCHR2(G1009A/K1012A)-FM* #7 lines with *chr2-1*$^{+/-}$*;PMIR159a-FM-GUS* #3 and *chr2-1*$^{+/-}$*;PMIR159b-FM-GUS* #5 lines, respectively. In the F2 generations, homozygous double mutants were identified by PCR using primers listed in Supplementary Table 4.

Plants were grown as previously described[39]. Wild-type (Col-0), mutants and transgenic lines were grown under a 12 h light–12 h dark cycle. Typically, three-week-old plants were harvested for various assays including picture-taking, sRNA blot and western blot analyses unless specifically mentioned. No statistical methods were used to predetermine sample size. Randomizations and blinding design were not relevant to this study.

Vector construction. Most coding DNA sequences (CDSs) and genomic sequences were cloned into pENTR/D-TOPO (Thermo Fisher) vectors and confirmed by sequencing. The majority of plant binary constructs were made using the Gateway system (Thermo Fisher). The primers for all constructs are listed in Supplementary Table 4.

PCHR2-gCHR2-FM-3′UTR and its derivatives were constructed as follows: first, three truncations of *CHR2* genomic fragment were amplified and cloned into pENTR/D vectors to generate pENTR-gCHR2-Truncation1–3; then, pENTR-gCHR2-Truncation1 and 2 were digested with XhoI/AscI and ligated to create pENTR/D-gCHR2-Truncation1+2; later, pENTR-gCHR2-Truncation1+2 and Truncation 3 were digested with BamHI/AscI, and ligated to create pENTR/D-gCHR2. Second, a fragment of Flag–4Myc (FM) was amplified and inserted 3′ of the DC cassette in a pBA002a-FM-DC vector[12] to generate pBA002a-FM-DC-FM. Subsequently, the 3′ UTR of *CHR2* (493 bp) was amplified and cloned into the SpeI/SacI-digested pBA002a-FM-DC-FM vector to yield pBA002a-FM-DC-FM-3′UTR. In parallel, the promoter of *CHR2* (1,903 bp) was amplified and cloned into the XbaI/EcoRV-digested pBA002a-FM-DC to yield pBA002a-PCHR2-FM-DC. The resultant pBA002a-PCHR2-FM-DC vector was digested with ApaI/XbaI and ligated to the same enzyme-treated pBA002a-FM-DC-FM-3′UTR to produce pBA002a-PCHR2-FM-DC-FM-3′UTR. The plasmid was digested with XbaI/AscI to remove the FM tag in front of the DC cassette, blunted by Klenow treatment, and self-ligated to create the pBA002a-PCHR2-DC-FM-3′UTR vector. Finally, the full length of the *CHR2* genomic fragment from a pENTR vector was transferred into pBA002a- PCHR2-DC-FM-3′UTR by Gateway attL-attR (LR) recombination reaction (ThermoFisher) to yield pBA002a-PCHR2-gCHR2-FM-3′UTR.

Site-specific mutations were introduced into *CHR2* by PCR using pENTR-gCHR2 truncations as templates. Then the fragments containing point mutations were swapped into pENTR-gCHR2 using restriction enzyme digestion followed by ligation. Finally, the *CHR2* genomic fragment containing site-specific point mutations was transferred into pBA002a-PCHR2-DC-FM-3′UTR by LR reaction to yield pBA002a-PCHR2-gCHR2 mutants-FM-3′UTR.

PMIR-FM-GUS were constructed as follows: First, approximately 1.8–2.4-kb promoters of *MIR159a*, *MIR159b*, *MIR164a*, *MIR164b* and *MIR164c* were amplified using pairs of primers listed in Supplementary Table 4, and cloned into pENTR/D vectors. Second, pBA002a-FM-DC-FM (as above) was treated with XbaI/AscI to remove the FM tag in front of the DC cassette, blunted by Klenow treatment, and self-ligated to produce the pBA002a-DC-FM vector; then, *GUS* CDS was amplified from pBA-GUS and cloned into the XhoI/PacI-digested pBA002a-DC-FM to create pBA002a-DC-FM-GUS. Finally, the promoters of MIR genes were transferred into pBA002a-DC-FM-GUS to yield PMIR-FM-GUS vectors.

P35S-FM-SE and PSE-FM-SE were constructed as follows: For P35S-FM-SE, *SE* CDS was transferred into pBA-FM-DC[12] by LR reaction to yield he P35S-FM-SE construct. For PSE-FM-SE, the *SE* promoter (2,241 bp) was amplified and cloned into BamHI/XbaI-digested pBA002a-FM-DC to yield pBA002a-PSE-FM-DC. Then the *SE* CDS was transferred into pBA002a- PSE-FM-DC by LR reaction to yield PSE-FM-SE vectors.

P35S-CHR2-YFP and P35S-SE-CFP were constructed as follows: For P35S-CHR2-YFP, first, two truncations of *CHR2* CDS were amplified and cloned into pCR-BluntII-TOPO vectors. Next, NotI/NdeI-digested *CHR2* truncation 1 and NdeI/AscI-digested *CHR2* truncation 2 were ligated into NotI/AscI-digested pENTR/D to yield pENTR/D-CHR2. Then, *CHR2* CDS was transferred into pBA-DC-YFP by LR reaction to create a P35S-CHR2-YFP vector. For P35S-SE-CFP, the *SE* CDS was transferred into pBA-DC-CFP by LR reaction to create P35S-SE-CFP vectors.

GST-CHR2-957-1823aa-6×His and its variants were constructed as follow: CHR2-957-1823aa-6×His and its derivatives were amplified from pENTR-CHR2 and pENTR-CHR2 mutants using sets of primers: 5′ forward containing an EcoRI digesting site and 3′ reverse primer containing 6×His tag sequence and SalI digestion site, respectively (Supplementary Table 4). Then the PCR products were cloned into an EcoRI/SalI-digested pGEX-4T-1 vector to produce pGEX-4T-1-GST-CHR2-957-1823aa-6×His and pGEX-4T-1-GST-CHR2-957-1823aa derivatives-6×His.

GST-6×His-DCL1, 6×His-SUMO-HYL1 and 6×His-SUMO-SE were constructed as follows: For the GST-6×His-DCL1 construct, pBA002a-Flag-DCL1 was digested with XbaI/SpeI and blunted with Klenow treatment. Next, the *Flag-DCL1* blunt fragment was ligated into SmaI-digested pAcGHLT-C (BD Biosciences) to yield the pAcGHLT-GST-6×His-DCL1 construct. For the 6×His-SUMO-SE construct, *SE* CDS was amplified and ligated into a BamHI/XbaI-digested pET28a-Avi-6×His-SUMO vector[40] to obtain the pET28a-Avi-6×His-SUMO-SE construct. For the 6×His-SUMO-HYL1 construct, *HYL1* CDS was amplified from the pBA-Myc-HYL1 plasmid[12] and digested with BamHI/SacI. The resultant fragment was ligated into the BamHI/SacI-treated pET28a-6×His-SUMO vector[40] to produce the pET28a-6×His-SUMO-HYL1 construct.

Yeast-two-hybrid (Y2H) and three-hybrid (Y3H) vectors were constructed as follows: For Y2H vectors, full-length and truncation derivatives of *SE*, *DCL1*, *HYL1* and *CHR2* or *CHR2* variants were cloned into vectors of pGADT7-GW, which harbours a GAL4 activation domain (AD), and pGBKT7-GW, which harbours a DNA binding domain (BD), using the LR reaction. Similarly, full-length *AtPrp40b* CDS was amplified and cloned into a pENTR/D-TOPO vector and confirmed by sequencing. Then the *AtPrp40b* CDS was cloned into a pGADT7-GW vector using the LR reaction. For Y3H vectors, a synthesized DNA fragment containing NotI, KpnI, SpeI, NcoI, AscI and BglII digestion sites was digested by NotI and BglII and cloned into a NotI/BglII-digested pBridge (Clontech) vector to produce pBridge-MCSII. Then the *SE* CDS amplified from pENTR/D-SE was cloned into the pBridge-MCSII vector by EcoRI/BamHI to yield pBridge-BD-SE-MCSII. Finally, the *CHR2* CDS fragment was obtained from pENTR-CHR2 using NotI/AcsI digestion and cloned into the NotI/AscI-digested pBridge-BD-SE-MCSII to producing the pBridge-BD-SE-CHR2 vector.

Gel filtration chromatography of protein extracts from plants. Nine-day-old *P35S-FM-SE* seedlings were harvested and homogenized in extraction buffer (20 mM Tris-HCl pH7.5, 150 mM NaCl, 4 mM $MgCl_2$, 75 μM $ZnCl_2$, 1% glycerol, 1 pellet per 12.5 ml Complete EDTA-free protease inhibitor (Roche), 1 mM PMSF, and 15 μM MG132). The homogenates were centrifuged twice at 15,000 r.p.m. for 15 min at 4 °C, and the final supernatant was filtered through a 0.2-μm filter. Next, the total protein extracts were injected into an AKTA FPLC system, and the proteins were fractioned through a Superdex 200 10/300 GL column (GE Healthcare). Fractions were collected for western blot analysis of FM-SE protein. The Superdex 200 column was also calibrated by gel filtration standard (Bio-Rad).

Immunoprecipitation and mass spectrometry. In brief, nine-day-old wild-type and *P35S-FM-SE* transgenic seedlings were harvested and ground in liquid nitrogen. Total proteins were extracted from 10 g ground powder in 40 ml immunoprecipitation buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 4 mM $MgCl_2$, 75 μM $ZnCl_2$, 1% glycerol, 1 pellet per 12.5 ml Complete EDTA-free protease inhibitor (Roche), 1 mM PMSF, and 15 μM MG132). After being cleared by ultracentrifugation, the protein complexes were immunoprecipitated using anti-FLAG M2 magnetic beads (Sigma, Cat#: M8823) at 4 °C for 2 h. After incubation, anti-FLAG M2 magnetic beads were washed four times with immunoprecipitation buffer for

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

5 min each at 4 °C, and protein complexes were then eluted by competitive 3×Flag peptide (100 μg/ml). The recovered protein complexes were re-suspended in the immunoprecipitation buffer and subsequently immunoprecipitated with anti-c-Myc-agarose affinity gel (Sigma, Cat#: A7470) at 4 °C for 1.5 h. Next, the beads were washed four times with immunoprecipitation buffer to remove unspecific-bound proteins. Finally, the protein complexes were eluted in elution buffer (5 mM EDTA, 200 mM NH_4OH) for mass spectrometry analysis in the Taplin Mass Spectrometry Facility at Harvard Medical School.

Y2H and Y3H assays. The Matchmaker Gold Y2H system (Clontech) was used for Y2H assays. The constructs were co-transformed into the Y2H Gold yeast strain (Clontech) and selected on medium lacking leucine, tryptophan, histidine and adenine for Y2H assays. The pBridge construct (Clontech) was used for Y3H assays. The constructs were co-transformed into the Y2H Gold yeast strain (Clontech) and selected on medium lacking leucine, tryptophan, histidine and methionine but supplemented with 5 mM 3-amino-1,2,4-triazole (3-AT) for Y3H assays. For both Y2H and Y3H assays, positive colonies were picked up and dropped on the selective plates with 1:5 (for Y2H) or 1:10 (for Y3H) serial dilutions for the picture recording.

Co-immunoprecipitation. Three-week-old plants on soil were sampled and ground in liquid nitrogen. Total proteins were extracted from 1 g ground powder in 5 ml immunoprecipitation buffer (40 mM Tris-HCl pH 7.5, 50 mM KCl, 5 mM $MgCl_2$, 5 mM DTT, 0.2 mM EDTA, 0.2% Triton X-100, 2% glycerol, 1 mM PMSF, 25 μM MG132, 1 pellet per 10 ml Complete EDTA-free protease inhibitor (Roche)); then, protein extracts were immunoprecipitated with the anti-FLAG M2 magnetic bead at 4 °C for 2 h; for RNase treatment, 250 μl RNase A (1 mg/ml) was added to 5 ml immunoprecipitation buffer during incubation. After incubation, the beads were washed four times with immunoprecipitation buffer at 4 °C for 5 min before application of SDS loading buffer at 95 °C for 10 min.

Confocal microscopy. Protoplast preparation and transfection were performed as described[41]. Transfected *Arabidopsis thaliana* protoplasts were imaged on an Olympus FV1000 confocal microscope.

Luciferase complementary imaging assay (LCI assay). The CDSs of *CHR2* and *SE* were cloned into *pCAMBIA-NLuc* and *pCAMBIA-CLuc*, respectively, by the LR reaction. Then the constructs were transformed into *A. tumefaciens* strain GV3101. LCI assays were performed as described[42].

sRNA and western blot analyses. sRNA and western blot assays were performed as described[39]. The sequences of the oligo probes for sRNA blot assays are listed in Supplementary Table 4. Western blot assays of CHR2-FM and its variants were typically conducted with an anti-Myc antibody throughout the study (Sigma, Cat#: C3956). Western blot assays of FM-GUS protein were performed with an anti-Flag antibody (Sigma, Cat#: F1804). Other endogenous protein-specific antibodies used in the study included anti-actin (Sigma, Cat#: A0480), anti-HYL1 (from B. Yu's laboratory), anti-DCL1 (from B. Yu's laboratory) and anti-SE (Agrisera, Cat#: AS09 532A). Secondary antibodies were goat-developed anti-rabbit and anti-mouse IgG (GE Healthcare, Cat#: NA934 and NA931).

sRNA sequencing and bioinformatics. Total RNA was prepared from three-week-old plants grown on soil using TRI Reagent (Sigma). Construction of sRNA libraries, Illumina sequencing and bioinformatic analysis were performed as described[39]. Until now, 371 miRNAs have been annotated in *Arabidopsis*, and we initially calculated their expression levels (reads per million, RPM) based on total reads. This analysis revealed that 33 miRNAs were significantly over-accumulated in the *chr2-1* mutant. However, the RPM values of numerous miRNAs including the founding members (miR159 and miR166) did not match the results that were consistently obtained from sRNA blot analysis throughout our experiments. This observation suggested that high-throughput sequencing using standard barcodes might introduce a bias against many species of miRNAs as reported previously[43,44]. To solve this issue, we normalized expression reads of miRNAs based on miR390, which consistently showed no changes between *chr2-1* and Col-0 plants in both sRNA blots and sRNA sequencing assays.

GUS histochemical analysis. Three-week-old transgenic plants were sampled and placed into a sodium phosphate solution containing 2 mM X-Gluc. The plants were vacuum infiltrated for 45 min and incubated at 37 °C overnight. After removal of chlorophyll, pictures of plants were taken with an Olympus SZH10 stereo microscope.

RNA extraction, reverse transcription, quantitative real-time PCR (RT–qPCR), and semiquantitative PCR. Total RNA was extracted from the three-week-old plants on soil using TRI Reagent (Sigma). cDNA synthesis, quantitative real-time PCR and semiquantitative PCR were performed as previously described[39]. Primers are listed in Supplementary Table 4.

Alternative splicing analysis. Alternative splicing in *chr2-1* was examined using rMATS software[45]

Gene ontology analysis. Gene ontology analyses were performed with the agriGO toolkit[46].

Expression and purification of recombinant proteins. GST-6×His-DCL1 protein was expressed in a baculovirus/insect cell expression system, whereas 6×His-SUMO-HYL1, 6×His-SUMO-SE, GST-CHR2-957-1823aa-6×His and its derivatives were expressed in *Escherichia coil* BL21 (DE3) cells. All protein purification was performed at 4 °C, and purified proteins were finally frozen by liquid nitrogen and stored at −80 °C.

For expression and purification of DCL1, pAcGHLT-GST-6×His-DCL1 was co-transfected with BaculoGold baculovirus DNA (BD Biosciences, Cat # 554740) into sf9 insect cells (BD Biosciences Cat# 554738; authenticated by the vendor BD Biosciences) to generate recombinant baculovirus according to the manufacturer's instructions. The recombinant viruses were amplified for two rounds, and P3 virus was collected for large-scale protein expression. P3 virus was added to 2.5×10^6 sf9 insect cells per ml for propagation, and insect cells were collected 62 h later.

The DCL1 protein was purified by two-step affinity chromatography (Ni-NTA affinity and glutathione *S*-transferase affinity) followed by gel filtration chromatography. The cell pellet was re-suspended in lysis buffer (100 mM Tris-HCl pH 8.0, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 1% Triton X-100, 1 pellet per 50 ml Complete EDTA-free protease inhibitor (Roche)) and disrupted with a high pressure homogenizer (AVESTIN, Cat#: EF-C3). After centrifugation and filtration with 0.4-μm membrane, the cleared lysate was supplemented with 20 mM imidazole and loaded on a HisTrap HP column (GE Healthcare, Cat#: 17-5248-02). The column was washed with 25 ml wash buffer (40 mM Tris-HCl pH 8.0, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 80 mM imidazole) and eluted with gradient elution buffer from 80 to 150 mM imidazole (40 mM Tris-HCl pH 8.0, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF). The peak fractions containing recombinant GST-6×His-DCL1 proteins were pooled and dialysed with 2 litre GST dialysis buffer (40 mM Tris-HCl pH 7.5, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 1 mM DTT) at 4 °C for 4 h. Then the dialysed fractions were supplemented with 1% Triton X-100 and loaded on a GSTrap HP column (GE Healthcare, Cat#: 17-5282-01). The column was washed with 25 ml wash buffer (40 mM Tris-HCl pH 7.5, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 1 mM DTT) and eluted with gradient elution buffer from 0 mM to 15 mM reduced glutathione (40 mM Tris-HCl pH 8.0, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 1 mM DTT). The peak fractions were collected and treated with thrombin (Calbiochem, Cat#: 605157-1KU) at 4 °C overnight to remove the GST-6×His tag. The fractions were concentrated by 50 kDa molecular weight cut-off (MWCO) centricon (Millipore, Cat#: UFC905024), and loaded to HiLoad 26/600 Superdex 200 pg column (GE Healthcare). The gel filtration buffer contained 40 mM Tris-HCl pH 7.5, 300 mM KCl, 5 mM β-mercaptoethanol and 2 mM DTT. The peak fractions containing DCL1 were collected and dialysed with one litre dialysis buffer (20 mM Tris-HCl buffer pH 7.5, 40 mM KCl, 2 mM β-mercaptoethanol, 2 mM DTT, 50% glycerol) 4 °C for 6 h. The final purified protein was quantified by SDS PAGE and aliquoted for storage at −80 °C.

For expression of recombinant proteins in *E.coli.*, transformed BL21 DE3 cells were grown in Terrific Broth (TB) for recombinant CHR2 and its variants or in Luria Broth (LB) for recombinant SE and HYL1 proteins. Cells were grown at 37 °C until $OD_{600nm} = 0.6$. Expression of recombinant proteins was typically induced with 0.5 mM IPTG at 16 °C overnight.

For purification of HYL1, the induced bacterial cells were collected and re-suspended in lysis buffer (40 mM Tris-HCl buffer pH 8.0, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 1% Triton X-100) and disrupted with a high pressure homogenizer (AVESTIN). After centrifugation and filtration with 0.4-μm membrane, the cleared lysate was supplemented with 20 mM imidazole and loaded onto a HisTrap HP column (GE Healthcare, Cat#: 17-5248-02). The column was washed with 25 ml wash buffer (40 mM Tris-HCl pH 8.0, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 80 mM imidazole) and eluted with gradient elution buffer from 80 to150 mM imidazole (40 mM Tris-HCl pH 8.0, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF). The peak fractions containing the recombinant 6×His-SUMO-HYL1 proteins were pooled and treated with SUMO protease at 4 °C overnight to remove the 6×His-SUMO tag. The fractions were concentrated by 50 kDa molecular weight cut-off (MWCO) centricon (Millipore), and loaded onto a HiLoad 26/600 Superdex 200 pg column (GE Healthcare). The gel filtration buffer contained 20 mM Tris-HCl pH 7.4, 50 mM KCl, 2 mM β-mercaptoethanol and 2 mM DTT. The peak fractions containing HYL1 were collected and concentrated again by 50 kDa molecular weight cut-off (MWCO) centricon (Millipore). The HYL1 protein was supplemented with 50% glycerol and finally frozen by liquid nitrogen and stored at −80 °C.

For purification of 6×His-SUMO-SE, the induced bacterial cells were collected and re-suspended in lysis buffer (20 mM Tris-HCl buffer pH 8.5, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 1% Triton X-100, 1 pellet per 50 ml Complete EDTA-free protease inhibitor (Roche)) and disrupted by the high pressure homogenizer (AVESTIN). After centrifuge and filtering (0.4 μm filter), the cleared lysate was supplemented with 20 mM imidazole and loaded on a HisTrap HP column (GE Healthcare, Cat#: 17-5248-02). The column was washed with 25 ml wash buffer (20 mM Tris-HCl pH 8.5, 300 mM KCl, 2% glycerol, 1 mM

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

β-mercaptoethanol, 1 mM PMSF, 80 mM imidazole) and eluted with gradient elution buffer from 80 to 150mM imidazole (20 mM Tris-HCl pH 8.5, 300 mM KCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF). The peak fractions containing the recombinant 6×His-SUMO-SE proteins were pooled and concentrated by 50 kDa molecular weight cut-off (MWCO) centricon (Millipore), and loaded onto a HiLoad 26/600 Superdex 200 pg column (GE Healthcare). The gel filtration buffer contained 20 mM Tris-HCl pH 8.5, 300 mM KCl and 5 mM β-mercaptoethanol. The peak fractions containing 6×His-SUMO-SE were collected and dialysed with one litre dialysis buffer (20 mM Tris-HCl buffer pH 8.5, 150 mM KCl, 5 mM β-mercaptoethanol, 50% glycerol) at 4 °C for 6 h. The final purified protein was quantified by SDS–PAGE and aliquoted for storage at −80 °C.

For purification of GST-CHR2-957-1823aa-6×His, the induced bacterial cells were collected and re-suspended in lysis buffer (50 mM sodium phosphate buffer pH 8.0, 500 mM NaCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 1% Triton X-100) and disrupted by the high pressure homogenizer (Avestin). After centrifugation and clearing with a 0.4-μm filter, the cleared lysate was supplemented with 20 mM imadazole and loaded onto a HisTrap HP column (GE Healthcare, Cat#: 17-5248-02). The column was washed with 25 ml wash buffer containing 80 mM imidazole, followed by 25 ml wash buffer containing 150 mM imidazole (50 mM sodium phosphate buffer pH 8.0, 500 mM NaCl, 2% glycerol; 1 mM β-mercaptoethanol, 1 mM PMSF). Then bound proteins were eluted with gradient elution buffer from 150 to 300 mM imidazole (50 mM sodium phosphate buffer pH 8.0, 500 mM NaCl, 2% glycerol; 1 mM β-mercaptoethanol, 1 mM PMSF). The peak fractions containing the recombinant GST-CHR2-957-1823aa-6×His proteins were pooled and dialysed with 2-l GST dialysis buffer (50 mM sodium phosphate buffer pH 7.6, 500 mM NaCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 1 mM DTT) at 4 °C for 4 h. Then fractions were supplemented with 1% Triton X-100 and loaded onto a GSTrap HP column (GE Healthcare, Cat#: 17-5282-01). The column was washed with 25 ml wash buffer (50 mM sodium phosphate buffer pH 7.6, 500 mM NaCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 1 mM DTT) and eluted with gradient elution buffer from 0 to 15 mM reduced glutathione (50 mM sodium phosphate buffer pH 8.0, 500 mM NaCl, 2% glycerol, 1 mM β-mercaptoethanol, 1 mM PMSF, 1 mM DTT). The peak fractions were collected and dialysed with one litre dialysis buffer (50 mM sodium phosphate buffer pH8.0, 500 mM NaCl, 5 mM β-mercaptoethanol, 2 mM DTT, 50% glycerol) at 4 °C for 6 h. The final purified protein was quantified by SDS–PAGE and aliquoted for storage at −80 °C. The same purification protocol was applied for purification of CHR2 derivatives.

ATPase assay. The method for the ATPase assay was modified from a previous description[27]. For Extended Data Fig. 2e, 1 pmol recombinant CHR2 or its derivatives were added into ATP hydrolysis reactions (20 μl) containing 20 mM sodium phosphate buffer pH 7.0, 25 mM KCl, 5 mM $MgCl_2$, 2 mM DTT, 500 ng dsDNA, 0.1 μl γ-^{32}P ATP (PerkinElmer, 3,000 Ci/mmol). Note: the final pooled concentration of NaCl and KCl in the reaction systems was about 75 mM. Reactions were incubated at 37 °C for 2 h. For Fig. 5c and Extended Data Fig. 9g–i, the reactions were carried out in 20 mM Tris-HCl, pH 7.5, 5 mM $MgCl_2$, 2 mM DTT, 0.3% NP-40, 1 U/μl SUPERase-In RNase Inhibitor (for pri-miRNA stimulated ATPase assay, Thermo Fisher). The final pooled concentration of NaCl and KCl in the reaction systems was ~60 mM, of which 50 mM was from CHR2 protein dialysis buffer and 10 mM was from dsDNA or pri-miRNA-dissolved buffer. 40 nM recombinant CHR2 or its derivatives was pre-incubated with dsDNA fragment (the same as in dsDNA EMSA) or with annealed pri-miR166f for 10 min before addition of 50 nM γ-^{32}P ATP and 100 nM cold ATP. Reactions were incubated at 37 °C for the indicated time points. Reactions were stopped by addition of 12.5 mM EDTA and 2.5 mM cold ATP and being placed on ice. For measurement of the substrate-regulated ATPase activity, 0.5 μM dsDNA or 1 μM pri-miR166f were added to the reactions. Liberated phosphate was analysed by thin layer chromatography (TLC) and phosphorimaging. The images were quantified with ImageQuant TL (GE Healthcare) and the ATP hydrolysis rates of CHR2 and its ATPase mutant were calculated using the data of 60 min reaction for Fig. 5c.

In vitro transcription and 5′ labelling of RNA/DNA. In vitro transcription and 5′ labelling of RNA/DNA substrates including pri-miR166f were performed as described[12]. The substrate of 5′ labelled dsDNA for EMSA is a PCR fragment containing the T7 promoter followed by the pri-miR166f sequence. The 5′ labelled ssDNA is a synthesized primer: 5′-$(A)_{97}$CCCTATAGTGAGTCGTATTA-3′. The substrate of 5′ labelled single-stranded (ss)RNA is an ssRNA (100 nt) in vitro transcribed using T7-G3A97_Rev primer[42] as a template.

In vitro Microprocessor assay. We added 1 pmol recombinant DCL1, 2 pmol HYL1, 2 pmol SE, 0.5 pmol CHR2 or its derivatives and 1,000 counts per minute (c.p.m.) of RNA substrate to 30 μl assay buffer containing 20 mM Tris-HCl pH7.5, 50 mM KCl, 4 mM $MgCl_2$, 1 mM DTT, 5 mM ATP, 1 mM GTP and 1 U/μl SUPERase-In RNase Inhibitor (Thermo Fisher). The final pooled concentration of NaCl and KCl was ~70 mM, of which ~20 mM salt was from protein dialysis buffer and RNA-dissolved buffer. The reconstitution assay was carried out at 37 °C. The reactions were stopped by adding 1 volume TBE-Urea sample buffer (Bio-Rad), being heated at 95 °C for 10 min and then being chilled on ice. The DCL1-processed products were fractionated using 10% denaturing polyacrylamide gel and detected overnight with a phosphor imaging plate (GE Healthcare). The images were quantified with Gel-Pro Analyzer (Media Cybernetics) software.

Electrophoretic mobility shift assays. Recombinant proteins and labelled RNA or DNA were mixed in the EMSA buffer (20 mM Tris-HCl pH 7.5, 2 mM $MgCl_2$, 2 mM DTT, 5 mM ATP, 0.3% NP-40, 1 U/μl SUPERase-In RNase Inhibitor (for RNA EMSA, Thermo Fisher)). The final pooled concentration of NaCl and KCl was ~55 mM, of which 50 mM was from protein dialysis buffer and 5 mM from DNA- or RNA-dissolved buffer. Mixtures were incubated at room temperature for 30 min. Bound complexes were resolved on native 1% agarose gel and visualized by radiography. The images were quantified with Gel-Pro Analyzer (Media Cybernetics) software. The K_d for SE and HYL1 and apparent K_d for CHR2 and its variants were calculated using Prism 5 (GraphPad) software. For the modelling prediction for the CHR2-nucleic acid binding pattern, the original data were analysed with GraphPad Prism 5[6]; specific binding with a Hill slope model was the best fit.

UV-crosslinking and ribonucleoprotein immunoprecipitation. Three-week-old plants grown on soil were irradiated four times with UV at 150 mJ/cm^2. The fixed samples were ground in liquid nitrogen and nuclei were isolated using a chromatin immunoprecipitation (ChIP) protocol as described[47]. The nuclei extracts were diluted with RIP buffer (40 mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 5 mM DTT, 0.2% Triton X-100, 2% glycerol, 1 mM PMSF, 25 μM MG132, 1 pellet per 10 ml Complete EDTA-free protease inhibitor (Roche) with 10 U/ml TURBO DNase (Thermo Fisher)) and immunoprecipitated with an anti-c-Myc agarose (Sigma, Cat#: A7470) at 4 °C for 4 h. Immunoprecipitates were washed three times with RIP buffer and once with high salt buffer (40 mM Tris-HCl pH 7.5, 500 mM KCl, 5 mM $MgCl_2$, 5 mM DTT, 0.2% Triton X-100, 2% glycerol, 1 mM PMSF, 25 μM MG132, 1 pellet per 10 ml Complete EDTA-free protease inhibitor (Roche)) at 4 °C for 5 min, followed by two washes with the proteinase K buffer (100 mM Tris-HCl pH 7.5, 50 mM NaCl, 10 mM EDTA). The beads were treated with Proteinase K (4 mg/ml) in 150 μl Proteinase K buffer at 37 °C for 20 min. After treatment with Proteinase K, the RNA was extracted using TRI Reagent (Sigma) and treated with TURBO DNase (Thermo Fisher). One half of the RNA was directly used for RT–qPCR. The other half (as a control) was further treated with RNase A (Sigma) and RNase T1 (Thermo Fisher) before RT–qPCR.

dsDNA and pri-miRNA remodelling/unwinding assays. For the pri-miRNA substrate, a 5′ ^{32}P labelled 26-nt RNA fragment was annealed to a truncated strand of pri-miR166f to generate a nicked pri-miR166f (Extended Data Fig. 9b). For the dsDNA substrate, a 5′ ^{32}P labelled 19-nt ssDNA fragment was annealed to a long ssDNA fragment to generate a dsDNA with long 5′ overhang (Extended Data Fig. 9e). The sequences of primers used to make dsDNA or pri-miRNA are listed in Supplementary Table 4. Recombinant CHR2 or its derivatives and annealed dsDNA or pri-miRNA were mixed in the remodelling buffer (20 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$, 2 mM DTT, 10 mM ATP, 1 mM GTP, 0.3% NP-40, 1 U/μl SUPERase-In RNase Inhibitor (specific for the pri-miRNA remodelling assay)). The final pooled concentration of NaCl and KCl was ~55 mM, of which 50 mM was from CHR2 protein dialysis buffer and 5 mM from dsDNA- or RNA-dissolved buffer. Before protease K treatment, the reactions were incubated at 37 °C (for DNA) or 25 °C (for RNA) for the indicated times. The ssDNA and dsDNA or RNA were fractionated using 15% (for DNA) and 12% (for RNA) native PAGE. The signals were detected by radiography. The images were quantified with ImageQuant TL (GE Healthcare).

In vivo DMS modification. For primer extension assays to probe in vivo RNA folding, approximately 3 g three-week-old *hyl1-2 chr2-1;PCHR2-gCHR2-FM* and *hyl1-2 chr2-1a* plants grown on soil were collected and completely covered in 20 ml 1× DMS reaction buffer (40 mM HEPES pH 7.5, 100 mM KCl and 0.5 mM $MgCl_2$) in a 50-ml Corning tube. DMS (Sigma, Cat#: D186309) was added to a final concentration of 0.75% as described[28]. Mock treatment was performed by addition of deionized water. Samples were treated in the DMS reaction buffer or mock solution at room temperature under a vacuum condition. Different time courses were initially tested; finally, a period of 30 min was found to be the optimal incubation time for adult plants (Extended Data Fig. 11c). To quench the reaction, 5 ml β-mercaptoethanol was added to a final concentration of 20%, and the mixture was incubated for 2 min under vacuum. After washing 4 times with 50 ml deionized water, the samples were immediately frozen with liquid N_2 and ground into powder.

For target-specific DMS–MaPseq, three-week-old Col-0 and *chr2-1* plants grown on soil were collected and treated with or without DMS using the same conditions as above except for DMS concentration. Several dosages of DMS were also tested; eventually 1% DMS was chosen for the assay because plant materials turned brown under treatment with 5% and 1.5% DMS and total RNA appeared to decay under these conditions.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Nuclear RNA extraction. Nuclear RNA extraction was performed as described[48]. In brief, 0.3 g ground powder was dissolved in 5 ml lysis buffer (20 mM HEPES pH7.5, 20 mM KCl, 2.5 mM $MgCl_2$, 25% glycerol, 250 mM sucrose, 5 mM DTT, 1 U/10μl RNase inhibitor (Thermo Fisher) and 1× Proteinase inhibitor cocktail without EDTA (Roche)). The homogenate was filtered through a double layer of Miracloth. The flow-through was centrifuged at 1,500*g* for 10 min at 4 °C. After removal of supernatant, the pellet was washed twice with 5 ml nuclear resuspension buffer (20 mM HEPES pH 7.5, 2.5 mM $MgCl_2$, 25% glycerol, 0.2% Triton X-100, 20 μl RNase inhibitor (Thermo Fisher)). Then the total nuclear RNA was extracted using TRI Reagent (Sigma).

Primer extension assays for probing in vivo RNA folding. Probing of in vivo RNA secondary structure by primer extension assay was performed as described[28] with modifications. For each sample, 2 μg total nuclear RNA was treated with TURBO DNase (Thermo Fisher), followed by phenol–chloroform extraction. DNase-treated nuclear RNA was mixed with ~200,000 c.p.m. ^{32}P-radiolabelled gene-specific primer (Supplementary Table 4). The mixture was precipitated by ethanol and resuspended in 10 μl Tris-KCl solution (10 mM Tris-HCl pH 7.5 and 50 mM KCl). The solution was heated at 75 °C for 3 min, followed by incubation at 35 °C (for 18S rRNA) or 55 °C (pri-miR164b) for 15 min. 10 μl reverse transcription reaction buffer mixed with 1 mM DTT, 1 mM dNTPs, 1 μl RNase Inhibitor (Thermo Fisher) and 1 μl SuperScript III (for 18S rRNA) (Thermo Fisher) or SuperScript IV (for pri-miR164b) reverse transcriptase (Thermo Fisher) was added. The reaction proceeded for 1 h at 55 °C (for 18S rRNA) or 60 °C (for pri-miR164b). To stop the reaction and hydrolysation of RNA, 2 μl of 2 M NaOH was added and mixture was heated at 95 °C for 10 min. After neutralization by 5 M HCl, the mixture was phenol–chloroform extracted and precipitated. Then the cDNA was resuspended in loading buffer (95% deionized formamide, 0.025% bromophenol blue, 0.025% xylene cyanol, 5 mM EDTA and 0.025% SDS) and size fractionated on 8% denaturing polyacrylamide gel. Gel image was collected with a Typhoon FLA7000 (GE Healthcare) and bands were quantified using ImageQuant TL (GE Healthcare).

Target-specific DMS–MaPseq. Target-specific DMS–MaPseq was performed as described[29]. For each sample, 5 μg DNase-treated nuclear RNA was mixed with gene-specific RT primers (5 pmol each primer, up to 5 gene-specific primers in one reaction, Supplementary Table 4). The mixture was precipitated and resuspended in 10 μl Tris-KCl solution (10 mM Tris-HCl pH 7.5 and 50 mM KCl). The solution was heated at 75 °C for 3 min, followed by incubation at 57 °C for 15 min. Four microlitres 5× First-Strand buffer (250 mM Tris-HCl pH 8.3, 375 mM KCl, 15 mM $MgCl_2$), 1 μl 0.1 M DTT (prepared freshly), 1 μl RNase inhibitor (Thermo Fisher), 1 μl H_2O and 1 μl TGIRT-III (Ingex, Cat#: TGIRT50) were added. The mixture was then incubated at room temperature for 30 min. Then 2 μl 10 mM dNTP was added and the reverse transcription proceeded at 63 °C for 2.5 h. The reaction was stopped by adding 2 μl of 2.5 M NaOH and heating at 95 °C for 3 min. After neutralization by 5 M HCl, the mixture was phenol–chloroform extracted and applied to an illustra MicroSpin S-200 HR column (GE Healthcare) to remove RT primers and nucleotides. The cleaned cDNA was precipitated and resuspended in deionized water. Then pri-miRNAs were amplified using KOD hot start DNA polymerase (Millipore) with gene-specific primers (Supplementary Table 4). PCR bands were gel purified and normalized according to band intensity before library construction.

PCR products were mixed equally and fragmented into 50–200 bp using NEBNext dsDNA Fragmentase (NEB) following the manufacturer's protocol. After purification using QIAquick nucleotide removal kit (QIAGEN), the fragments were subjected to end repair, adenylation and adaptor ligation using Illumina adapters, mainly following the published protocol[49]. The fragments were barcoded through adaptor ligation. The ligation products were size fractionated on 3% low melting agarose gel, and 250–350 bp adaptor-ligated fragments were purified by gel excision. Next, the purified barcoded libraries were enriched by 12 cycles of PCR using KOD hot start DNA polymerase. Finally, the PCR products were cleaned using Agencourt AMPure XP beads (Beckman). The libraries were quantified using Agilent TapeStation before sequencing by 150 bp single-end reads on the Illumina NextSeq 500 at Texas A&M University.

Sequencing alignment and analysis. Raw fastq files were initially filtered for a quality control that requires a quality score >30 by sickle (https://github.com/ucdavis-bioinformatics/sickle) and trimmed to remove the adaptor sequences by Cutadapt[50]. Clean reads over 35 bp were retained. Reads were aligned using Burrows-Wheeler Aligner (BWA)[51] allowing 5% mismatches with the settings: bwa aln -n 0.05. The mapping results were sorted by Picard tools (http://broadinstitute.github.io/picard/) with coordination and the base calls of aligned reads to the reference sequence were summarized using samtools (v1.5) mpileup[52]. Nucleotide mismatches and sequencing depth were extracted using sequenza-utils (v2.1) pileup2acgt (https://pypi.python.org/pypi/sequenza-utils). The DMS signal was calculated as number of mismatches per sequencing depth for each adenine (A) and cytosine (C) nucleotide. The two biological replicates were normalized to the identical means of total DMS signal. Then the average DMS signal for each A and C nucleotide was calculated. Based on the DMS signal, the secondary structures of pri-miRNAs were modelled by RNAfold[53]. The threshold was used to separate adenine and cytosine bases into paired and unpaired nucleotides to produce the best fitting model for our experimental data. The threshold constraint varied between 0.01 to 0.03 depending on different pri-miRNAs. DMS signals were colour coded on structure models using VARNA[54] (http://varna.lri.fr,).

Quantification and statistical analysis. The images of Microprocessor activity and sRNA blot assays were quantified with Gel-Pro Analyzer (Media Cybernetics). The K_d for EMSA assays was calculated using Prism 5 (GraphPad) software. The images from primer extension assays, dsDNA and pri-miRNA remodelling/helicase assays and ATPase assays were quantified with ImageQuant TL (GE Healthcare). Unpaired, two-tailed Student's *t*-test was performed using the software Excel. All statistics are described below unless specifically mentioned in the Figures or Figure legends.

For Figs. 3, 5d, quantification of the Microprocessor cleavage efficiency was calculated by the ratio of processed to unprocessed pri-miR166f fragments. The relative efficiency was normalized to that of the Microprocessor reaction alone where the ratio was arbitrarily set at 1 with s.d. from three replicates.

For Fig. 4d, the relative signal of pri-miRNAs was normalized initially to the input and then to the control immunoprecipitate, where the ratio was arbitrarily set to 1 with s.d. from three biological repeats ($*P < 0.05$; $**P < 0.01$; unpaired, two-tailed Student's *t*-test).

For Fig. 5e, f and Extended Data Figs. 11e–h, 12, in the primer extension assays, the colour intensity of coded nucleotides along the pri-miR164b illustrates the relative DMS reactivity, calculated as the band signal intensity normalized to the highest level, which was set arbitrarily to 1.0. DMS (+/−) refers to the DMS treated and untreated samples, respectively. In the DMS–MaPseq analysis, DMS mutation frequencies of A and C residues (mismatch/total) from targeted pri-miRNAs averaged from two biological replicates were plotted along pri-miRNA sequence. Position 1 corresponds to the 5′ end of modelled regions. The overall and zoomed-in secondary structures of the targeted pri-miRNAs were modelled from DMS activity according to RNAfold[53]. The nucleotides that displayed different DMS activity on pri-miRNAs between Col-0 and *chr2-1* plants are colour-coded.

For Extended Data Figs. 4, previously published datasets[55–58] were reanalysed.

Graph drawing. Graphs with dot plots (individual data points) were drawn using GraphPad Prism 7.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

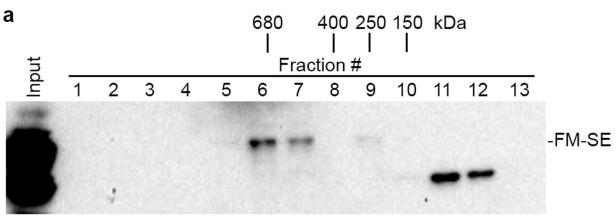
Data availability. Original data that support the findings of this study are available from the corresponding author upon reasonable request. The GEO accession numbers for sRNA sequence and DMS–MaPseq are GSE108858 and GSE108857, respectively. Source Data for graphs plotted in Figs. 1, 3–5 are available in the online version of this paper. Gel source data for Figs. 1, 2, 4, 5, Extended Data Figs. 1–5, 8–11 are available in Supplementary Fig. 1.

38. Zhang, X., Henriques, R., Lin, S. S., Niu, Q. W. & Chua, N. H. Agrobacterium-mediated transformation of *Arabidopsis thaliana* using the floral dip method. *Nat. Protocols* **1**, 641–646 (2006).
39. Zhu, H. et al. *Arabidopsis* Argonaute10 specifically sequesters miR166/165 to regulate shoot apical meristem development. *Cell* **145**, 242–256 (2011).
40. Zhao, B. et al. Structural basis for concerted recruitment and activation of IRF-3 by innate immune adaptor proteins. *Proc. Natl Acad. Sci. USA* **113**, E3403–E3412 (2016).
41. Yoo, S. D., Cho, Y. H. & Sheen, J. *Arabidopsis* mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat. Protocols* **2**, 1565–1572 (2007).
42. Zhang, Z. et al. RISC-interacting clearing 3′- 5′ exoribonucleases (RICEs) degrade uridylated cleavage fragments to maintain functional RISC in *Arabidopsis thaliana*. *eLife* **6**, e24466 (2017).
43. Raabe, C. A., Tang, T. H., Brosius, J. & Rozhdestvensky, T. S. Biases in small RNA deep sequencing data. *Nucleic Acids Res.* **42**, 1414–1426 (2014).
44. Sorefan, K. et al. Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* **3**, 4 (2012).
45. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).
46. Du, Z., Zhou, X., Ling, Y., Zhang, Z. & Su, Z. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* **38**, W64–W70 (2010).
47. Gendrel, A. V., Lippman, Z., Martienssen, R. & Colot, V. Profiling histone modification patterns in plants using genomic tiling microarrays. *Nat. Methods* **2**, 213–218 (2005).
48. Zhang, Z. et al. KETCH1 imports HYL1 to nucleus for miRNA biogenesis in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **114**, 4011–4016 (2017).
49. Feng, S., Rubbi, L., Jacobsen, S. E. & Pellegrini, M. Determining DNA methylation profiles using sequencing. *Methods Mol. Biol.* **733**, 223–238 (2011).
50. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

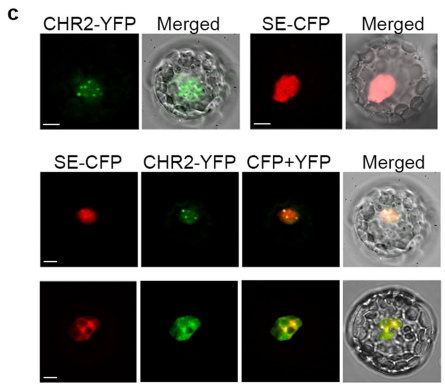
51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
52. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
54. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).
55. Li, C. et al. Concerted genomic targeting of H3K27 demethylase REF6 and chromatin-remodeling ATPase BRM in *Arabidopsis*. *Nat. Genet.* **48**, 687–693 (2016).
56. Bezhani, S. et al. Unique, shared, and redundant roles for the *Arabidopsis* SWI/SNF chromatin remodeling ATPases BRAHMA and SPLAYED. *Plant Cell* **19**, 403–416 (2007).
57. Tang, X. et al. The *Arabidopsis* BRAHMA chromatin-remodeling ATPase is involved in repression of seed maturation genes in leaves. *Plant Physiol.* **147**, 1143–1157 (2008).
58. Archacki, R. et al. BRAHMA ATPase of the SWI/SNF chromatin remodeling complex acts as a positive regulator of gibberellin-mediated responses in *Arabidopsis*. *PLoS One* **8**, e58588 (2013).
59. Allen, R. S. et al. Genetic analysis reveals functional redundancy and the major target genes of the *Arabidopsis* miR159 family. *Proc. Natl Acad. Sci. USA* **104**, 16371–16376 (2007).
60. Raman, S. et al. Interplay of miR164, CUP-SHAPED COTYLEDON genes and LATERAL SUPPRESSOR controls axillary meristem formation in *Arabidopsis thaliana*. *Plant J.* **55**, 65–76 (2008).
61. Archacki, R. et al. *Arabidopsis* SWI/SNF chromatin remodeling complex binds both promoters and terminators to regulate gene expression. *Nucleic Acids Res.* **45**, 3116–3129 (2017).
62. Laubinger, S. et al. Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **105**, 8795–8800 (2008).
63. Raczynska, K. D. et al. The SERRATE protein is involved in alternative splicing in *Arabidopsis thaliana*. *Nucleic Acids Res.* **42**, 1224–1244 (2014).
64. Knop, K. et al. Active 5' splice sites regulate the biogenesis efficiency of *Arabidopsis* microRNAs derived from intron-containing genes. *Nucleic Acids Res.* **45**, 2757–2775 (2017).
65. Manavella, P. A. et al. Fast-forward genetics identifies plant CPL phosphatases as regulators of miRNA processing factor HYL1. *Cell* **151**, 859–870 (2012).

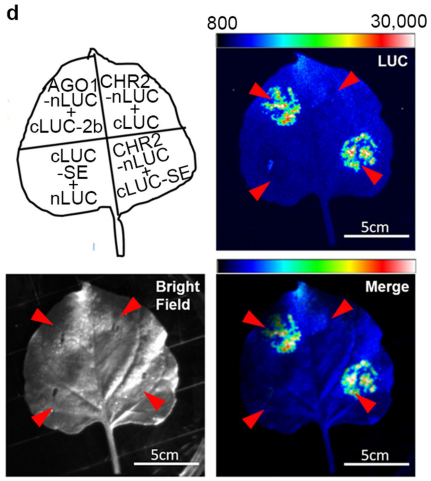
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

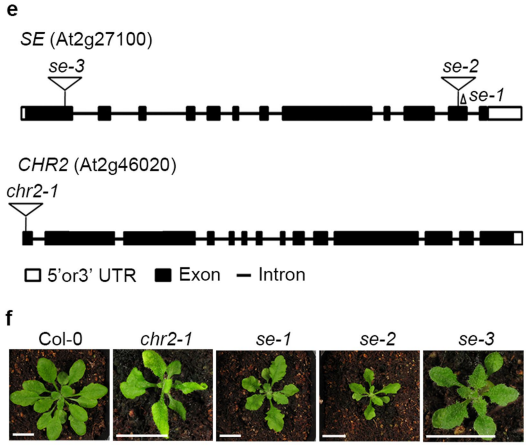


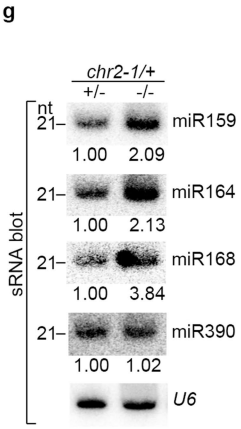
b

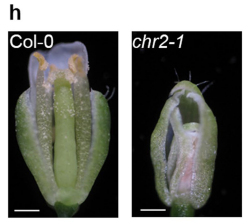
Position	Sequence	Scan.Charge	Score	ΔScore	Ions
359-372aa	MVNPAMNPFSGQGR R.M*VNPAM*NPFSGQGR.E	9,542.2 9,586.2	3.19 3.29	0.468 0.408	19/26 15/26
1574-1584aa	EVNATVADLSK R.EVNATVADLSK.K	5,664.2	2.70	0.377	16/20

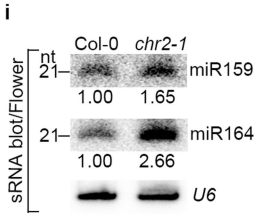


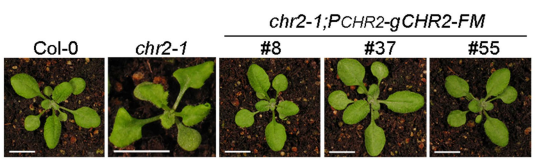


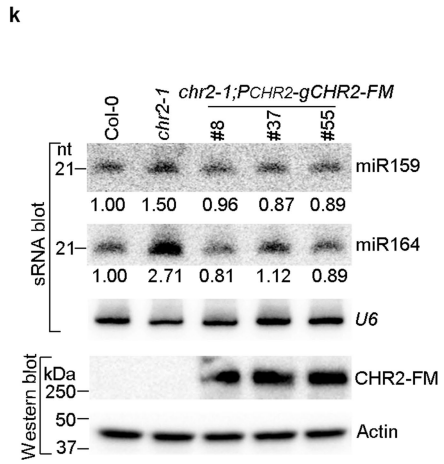


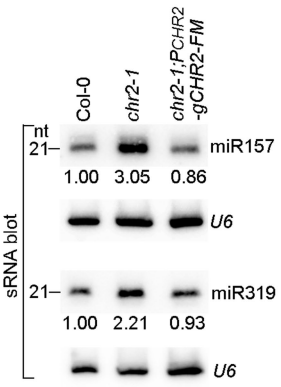


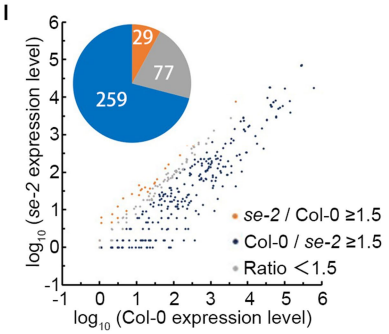


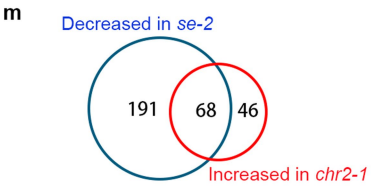










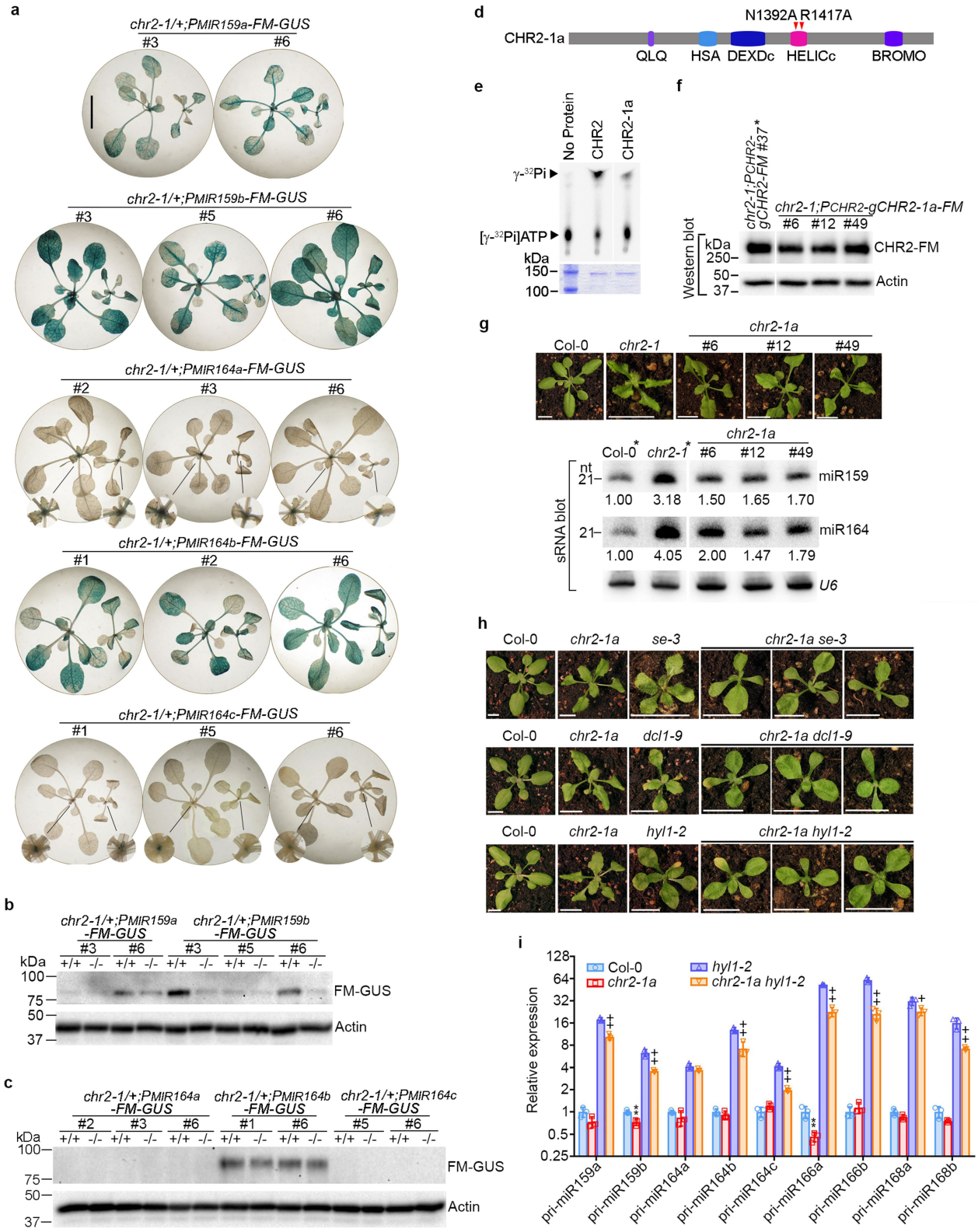


Extended Data Fig. 1 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 1 | CHR2 is a partner of the Microprocessor component SE and represses miRNA accumulation in *Arabidopsis*. **a**, Size-exclusion chromatography of total protein extracts from Col-0; *P35S- FM-SE* transgenic plants shows the FM-SE enrichment in macromolecular complexes (>680 kDa). Western blot analysis was conducted with an anti-Flag antibody. **b**, CHR2 peptides recovered from proteomic analysis of the SE complex but not a control immunoprecipitation from Col-0 plants. SE complexes were isolated from Col-0; *P35S-FM-SE* transgenic plants by two-step affinity purification, and analysed by mass spectrometry. **c**, Confocal microscopic images show co-localization of SE–CFP and CHR2–YFP when co-expressed in *Arabidopsis* protoplasts. Top, SE–CFP and CHR2–YFP, individually expressed, serve as control. Scale bar, 5 μm. **d**, Luciferase complementary imaging (LCI) assay shows the specific CHR2–SE interaction in *Nicotiana benthamiana*. The infiltration scheme in the leaf shows different combinations of constructs fused to either N-terminal (nLuc) or C-terminal (cLuc) regions of luciferase. LCI complementation (LUC), bright field, and merged photograph (Merge) are shown. The red arrows and colour bar indicate the infiltration positions and the signal intensity, respectively. **e**, Schematic illustration of T-DNA insertions (reverse triangles) or deletion (triangle) in *CHR2* and *SE* genes. **f**, Opposite leaf polarity phenotypes in *chr2-1* and *se* loss-of-function mutants. Pictures of 28-day-old plants are shown. Scale bar, 1 cm. **g**, Additional replicate of sRNA blot analysis shows enhanced accumulation of miRNAs in the *chr2-1* null mutant. **h**, Flower developmental defects in *chr2-1*. Scale bar, 0.5 mm. **i**, sRNA blot analysis shows that miRNAs accumulated in *chr2-1* flowers. **j**, The *chr2-1* morphological defect was fully rescued by the *PCHR2-gCHR2-FM* transgene. Scale bar, 1 cm. **k**, sRNA blot analysis shows that miRNA accumulation was restored to wild-type miRNA levels in the *chr2-1*; *PCHR2-gCHR2-FM* complementation lines (top). Western blot analysis of CHR2-FM using an anti-Myc antibody is shown (bottom). **l**, sRNA-seq revealed that the accumulation of 259 of 365 annotated miRNAs (http://www.mirbase.org) in *Arabidopsis* is SE-dependent. **m**, Overlap of SE-regulated and CHR2-repressed miRNAs in *Arabidopsis*.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

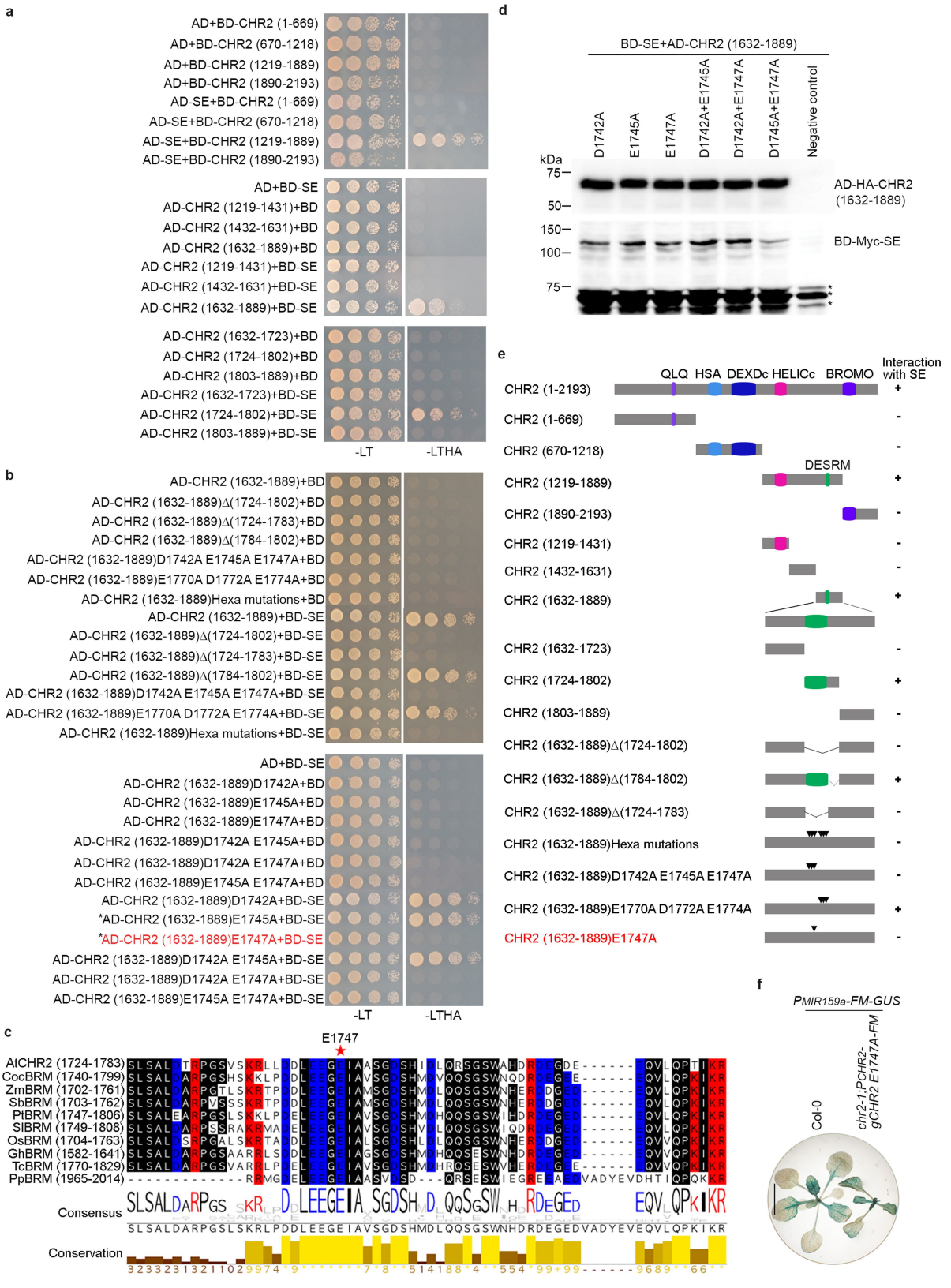


Extended Data Fig. 2 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 2 | CHR2 plays self-antagonistic roles in the regulation of miRNA accumulation at transcriptional and posttranscriptional levels. a, Histochemical staining analysis of GUS activity shows comparable transcriptional levels of tested MIR genes in wild-type (WT, left, big plants) and *chr2-1* (right, small plants) backgrounds. The promoters of individual MIR genes[59–61] were fused to the FM-GUS coding sequence and transformed into a *chr2-1*$^{+/-}$ heterozygote line. Twenty-one-day-old T2 generation plants were used for GUS staining. The SAM and leaf primordia of *PMIR164a-* and *PMIR164c-FM-GUS* transgenic plants are magnified to comparable sizes. Scale bar, 1 cm. **b**, **c**, Western blot analysis shows that protein levels of FM-GUS from the native MIR promoters were higher in wild-type (+/+) than in *chr2-1* (−/−) plants. Western blot assays were done with an anti-Flag antibody. Actin serves as a loading control. **d**, Schematic illustration of the conserved functional domains in CHR2 and the locations of the point mutations N1392A and R1417A in CHR2-1a (red arrows). **e**, Thin-layer chromatography (TLC) assay shows that the CHR2-1a variant has significantly reduced ATPase activity in vitro. ATPase activity was conducted with 500 ng dsDNA for 2 h. Bottom, SDS–PAGE of purified CHR2 and CHR2-1a proteins. **f**, Western blot analysis using an anti-Myc antibody shows comparable expression of CHR2 and CHR2-1a in the *chr2-1* background. Actin serves as a loading control. **g**, *chr2-1* morphological and miRNA accumulation defects, analysed by sRNA blot assay, were partially rescued in *chr2-1a* (*chr2-1;PCHR2-gCHR2-1a-FM*) plants. Scale bar, 1 cm. *U6* serves as a loading control for sRNA blot analysis. In **f**, **g**, *shows the same samples as in Extended Data Fig. 10e. **h**, Morphological phenotypes of the hypomorphic *chr2-1a* mutant and its derived double mutants. Photographs of 21-days-old plants are shown. Scale bar, 1 cm. **i**, RT–qPCR analysis of pri-miRNAs in single and double mutants relative to Col-0. Total RNA was prepared from a pool of 21-day-old plants; *EF-1α* serves as a loading control. The relative signals of pri-miRNAs were normalized to those of wild-type Col-0, where the ratio was arbitrarily set to 1 with s.d. calculated from three biological repeats. Individual data points are shown. Data significantly different from the corresponding controls are indicated (*chr2-1a* versus Col-0, $^{**}P < 0.01$; *chr2-1a hyl1-2* versus *hyl1-2*, $^{+}P < 0.05$, $^{++}P < 0.01$; unpaired, two-tailed Student's *t*-test).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

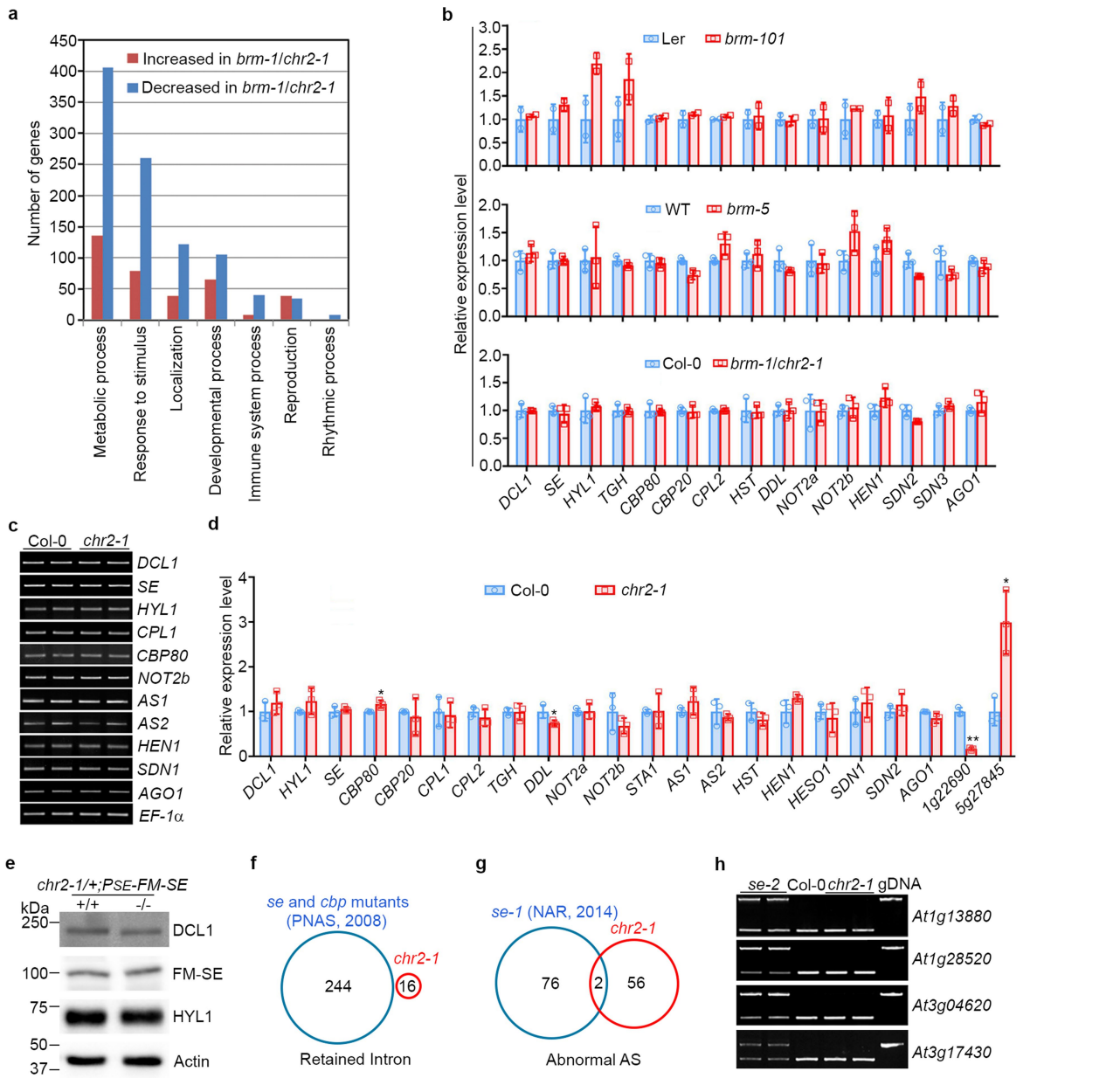


Extended Data Fig. 3 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

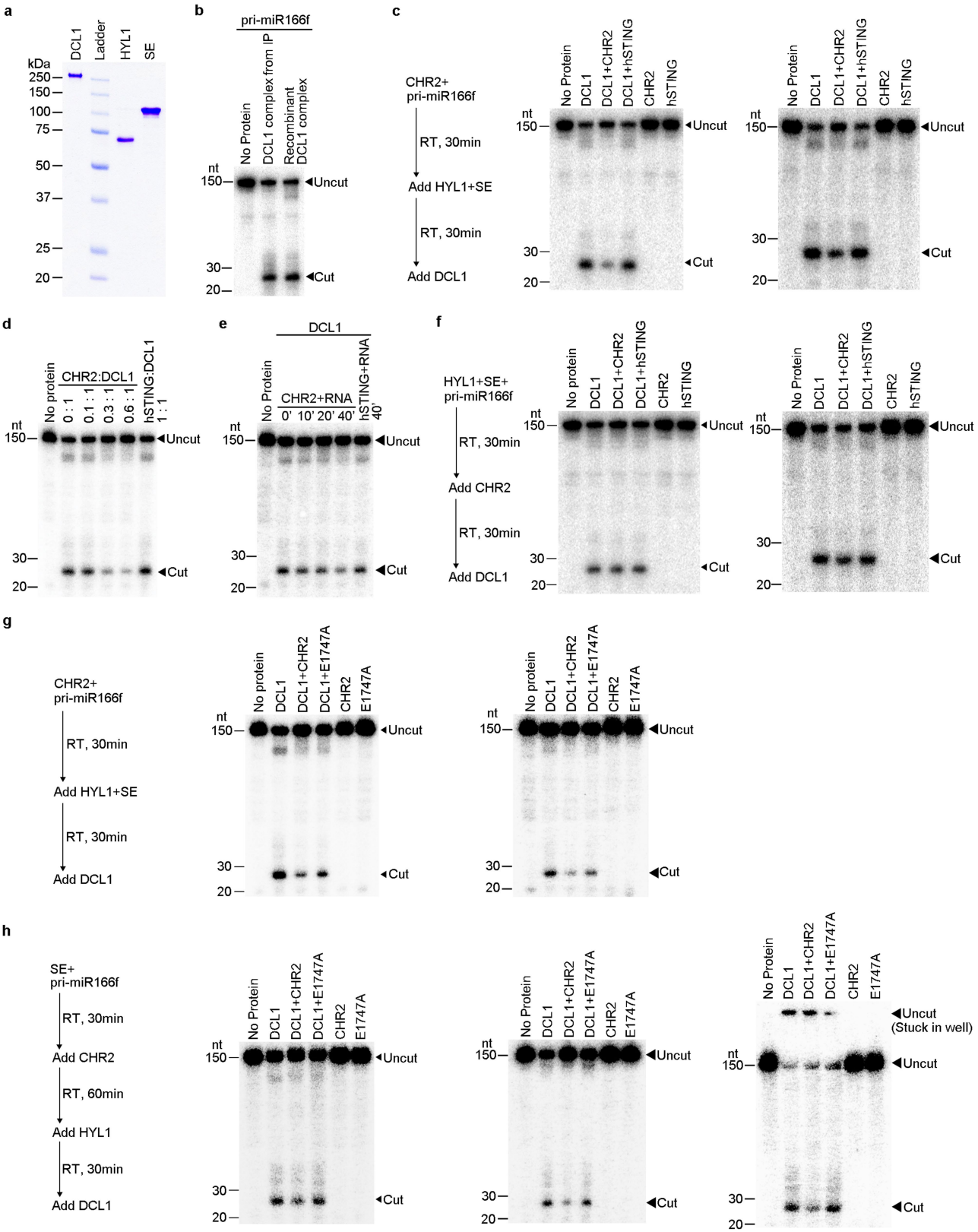
Extended Data Fig. 3 | Delineation of the CHR2–SE interaction interface and identification of CHR2 E1747 as a critical residue for the specific CHR2–SE interaction by Y2H assays. a, Rough mapping of CHR2 domains that interact with SE. **b**, Fine mapping of the CHR2–SE interaction interface and final identification of CHR2 E1747 as a critical residue for the CHR2–SE interaction. Note: part of the image related to CHR2 E1747A is also shown in Fig. 2a. AD, GAL4 activation domain; BD, GAL4 DNA binding domain. −LT, lacking Leu and Trp; −LTHA, lacking Leu, Trp, His and Ade. 1:5 serial dilutions are shown. **c**, The DESRM domain and E1747 residue of CHR2 that are involved in its interaction with SE are conserved through plants. Sequence alignment of DESRM domains of CHR2 proteins across different species (the region is rich in Asp-Glu-Ser residues, and therefore named the DES-rich motif (DESRM)). Red star indicates E1747, which is critical for the CHR2−SE interaction. Conserved neutral, acidic and basic amino acids are marked in black, blue, and red, respectively. **d**, Western blot analysis shows that expression levels of CHR2 variants were comparable in transformed yeast colonies. Western blot assays were conducted with anti-HA and anti-Myc antibodies to detect CHR2 and SE, respectively. Asterisks indicate non-specific bands. **e**, Summary of Y2H assays. Left, schematic illustration of CHR2 variants used for Y2H. The numbers indicate the positions of the amino acid residues in the constructs. Right, + and – denote positive and negative results for the CHR2−SE interaction from Y2H assays in **a** and **b**. **f**, Histochemical staining analysis of GUS activity from *PMIR159a-FM-GUS* homozygotes show that CHR2(E1747A) does not alter the transcriptional level of *MIR159a* in Col-0 and *chr2-1;PCHR2-gCHR2(E1747A)-FM* backgrounds. Scale bar, 1 cm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 4 | CHR2 does not affect the expression of canonical components involved in miRNA biogenesis and metabolism, nor the SE- dependent pre-mRNA splicing pathway. a, Gene Ontology analysis of differentially expressed genes in *chr2-1* plants. The dataset is mined from a previously published RNA-seq database[55] for the *chr2* mutant. **b**, Re-analysis of previously published microarray data sets for *chr2* mutants[56–58] to assess the expression of numerous canonical genes involved in the miRNA pathways. **c**, **d**, RT–PCR (**c**) and RT−qPCR (**d**) show that the expression levels of the genes involved in the miRNA pathway are generally comparable between Col-0 and the *chr2-1* mutant. *At1g22690* and *At5g27845* serve as controls for down- and upregulated genes in *chr2-1*. Data significantly different from Col-0 controls are indicated (* $P < 0.05$, ** $P < 0.01$; unpaired, two-tailed Student's *t*-test). In **b**, **d**, the relative signals of genes were normalized to those of wild-type, where the ratio was arbitrarily set to 1 with s.d. calculated from published data (**b**) or three biological repeats (**d**). Individual data points are shown. **e**, Western blot analysis shows that *chr2-1* does not show enhanced accumulation of Microprocessor proteins in vivo. Western blot assays were conducted using antibodies against native DCL1 and HYL1 proteins; an anti-Myc antibody was used to detect the native promoter-driven FM−SE protein. Actin serves as a control. **f**, **g**, Re-analysis of previously published RNA-seq data sets[55,62,63] for *chr2-1*/*brm-1* and *se* mutants shows that there is no significant overlap of abnormal intron retention events (**f**) and abnormal alternative splicing between *se* and *chr2-1* mutants (**g**). **h**, RT–PCR analysis of selected marker genes shows that CHR2 is not involved in the SE-coordinated pre-mRNA splicing pathway. gDNA (genomic DNA) serves as a control.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

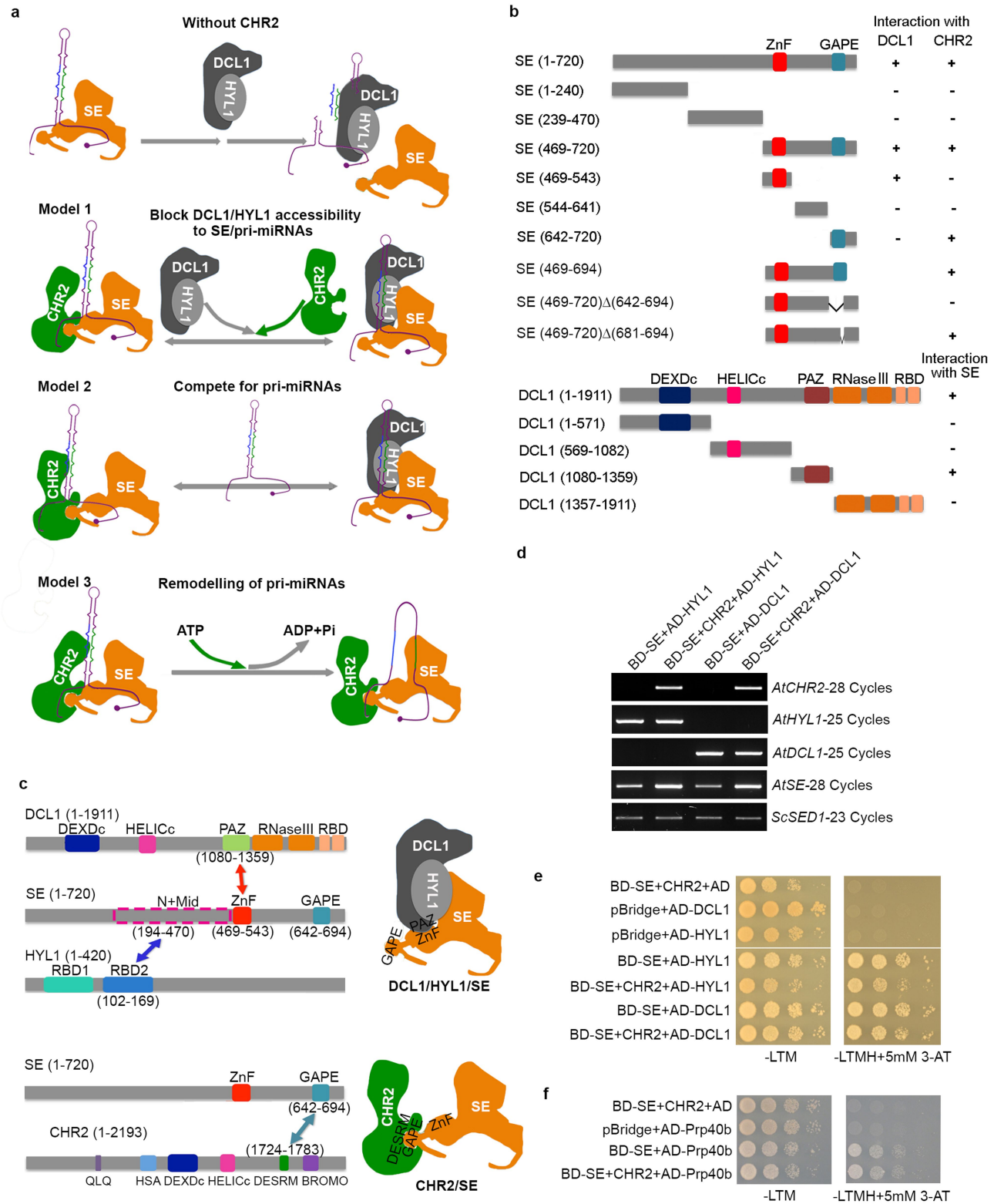


Extended Data Fig. 5 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 5 | In vitro Microprocessor assays show that CHR2 inhibits pri-miRNA processing. a, SDS–PAGE of recombinant proteins purified from a Baculovirus/insect system or from *E. coli*. **b**, The in vitro-reconstituted Microprocessor processed pri-miRNA substrates as accurately and efficiently as DCL1 complex immunoprecipitated from plants in vivo. **c**, Pre-incubation of CHR2 with pri-miRNAs inhibited the miRNA processing in vitro. hSTING serves as a negative control. On the left, schematic illustration of in vitro reconstitution orders with recombinant CHR2, 5′ ^{32}P-labelled pri-miR166f transcript and Microprocessor components. RT, room temperature. On the right, autoradiography of in vitro Microprocessor assays (two replicates). Black arrows indicate processed and unprocessed fragments of pri-miR166f. **d**, CHR2 inhibition of Microprocessor activity was positively correlated with its amount applied. Molar ratio CHR2:DCL1 is shown at the top of the autoradiograph. The reconstitution assay was conducted as in Fig. 3a. **e**, CHR2 inhibition of pri-miRNA processing was positively correlated with the pre-incubation time of CHR2 with pri-miRNA. The pre-incubation time of CHR2 with pri-miR166f is shown at the top of the autoradiograph. The reconstitution assay was conducted as in Fig. 3a. **f**, Incubation of HYL1−SE with pri-miRNAs before addition of CHR2 largely bypassed the inhibitory effect of CHR2 on pri-miRNA processing in vitro. Left, schematic representation of the in vitro reconstitution assay. Right, autoradiography of in vitro Microprocessor assays (two replicates). **g**, Incubation of CHR2 variant compromised in SE interaction (CHR2(E1747A)) with pri-miRNA substantially reduced pri-miRNA processing in vitro. Left, schematic representation of the in vitro reconstitution assay. Right, autoradiography of in vitro Microprocessor assays (two replicates). **h**, CHR2(E1747A) had only a marginal inhibitory effect on pri-miRNA processing when the substrate pri-miRNA was pre-incubated with excessive SE. Molar ratio of SE:CHR2 (or its variant) was 4:1. Left, schematic representation of in vitro reconstitution assay. Right, autoradiography of in vitro Microprocessor assays (three replicates).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

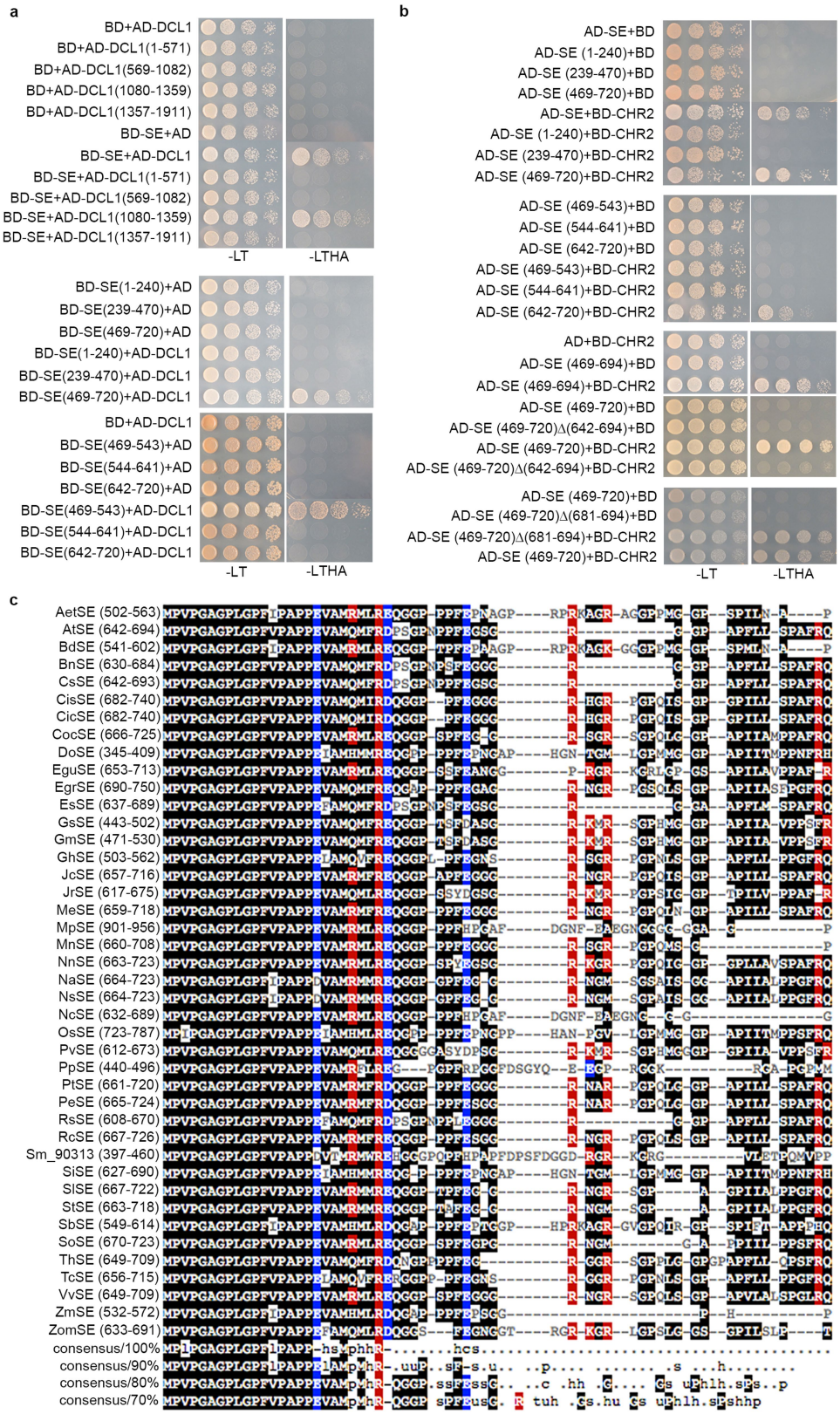


Extended Data Fig. 6 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

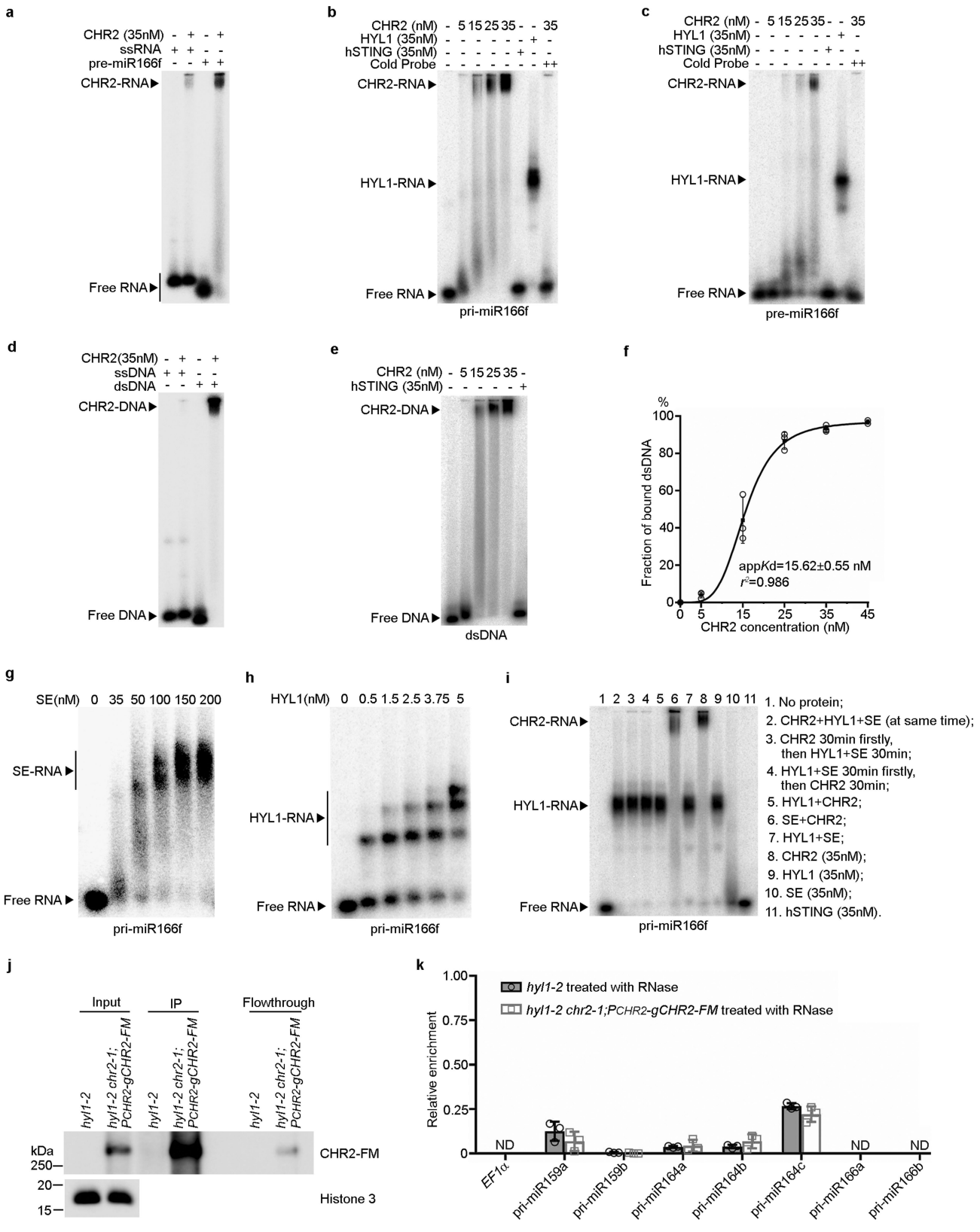
Extended Data Fig. 6 | Three non-exclusive hypotheses for the mechanism by which CHR2 inhibits miRNA production; and CHR2 and DCL1—HYL1 interact with different domains of SE. a, Schematic illustration of three non-exclusive hypotheses for the mechanism by which CHR2 inhibits miRNA processing. **b**, Summarized Y2H results for mapping of the DCL1—SE and CHR2—SE interaction interfaces in SE protein (detailed results are shown in Extended Data Fig 7a, b). Left, schematic illustration of the truncation and deletion variants of SE and DCL1 used for Y2H. The numbers indicate the positions of the amino acid residues in the constructs. Right, + and – denote positive and negative Y2H results for the interaction with SE, respectively. **c**, Proposed models of the protein interactions between DCL1—HYL1 and SE and between CHR2 and SE. DCL1 occupies the zinc finger domain (ZnF) of SE through its PAZ domain; dsRNA-binding domain 2 (RBD2) in HYL1 interacts with the N-terminal and mid domains (N+Mid) of the SE core[19]. The DESRM motif in CHR2 contacts the GAPE domain (named for enriched Gly-Ala-Pro-Glu residues) in the unstructured C terminus of SE. Numbers show the amino acid residues comprising the indicated domains. The red, blue and cyan arrows indicate the interaction domains of the indicated proteins. **d**, RT–PCR shows that all genes are expressed well in the Y3H system. All transfected yeast cells were cultured in —Leu, —Trp and —Met synthetic defined broth. *ScSED1* serves as a control. **e**, **f**, Y3H assays show that CHR2 does not interfere with the accessibility of DCL1—HYL1 (**e**) or that of U1 snRNP component (Prp40b)[64,65] (**f**) to SE protein. 1:10 serial dilutions are shown. —LTM, lacking Leu, Trp and Met; —LTMH+5mM 3-AT, lacking Leu, Trp, Met, His and adding 5 mM 3-AT.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | Mapping the regions of SE that interact with DCL1 and CHR2 by Y2H; and the GAPE domain of SE that interacts with CHR2 is conserved through plants. a, **b**, Detailed Y2H results of mapping the DCL1–SE interaction interface (**a**) and the interaction region of SE with CHR2 (**b**); 1:5 serial dilutions are shown. −LT, lacking Leu and Trp; −LTHA, lacking Leu, Trp, His and Ade. **c**, Sequence alignment of GAPE domains of SE proteins across different species; conserved neutral, acidic and basic amino acids are marked by black, blue, and red, respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

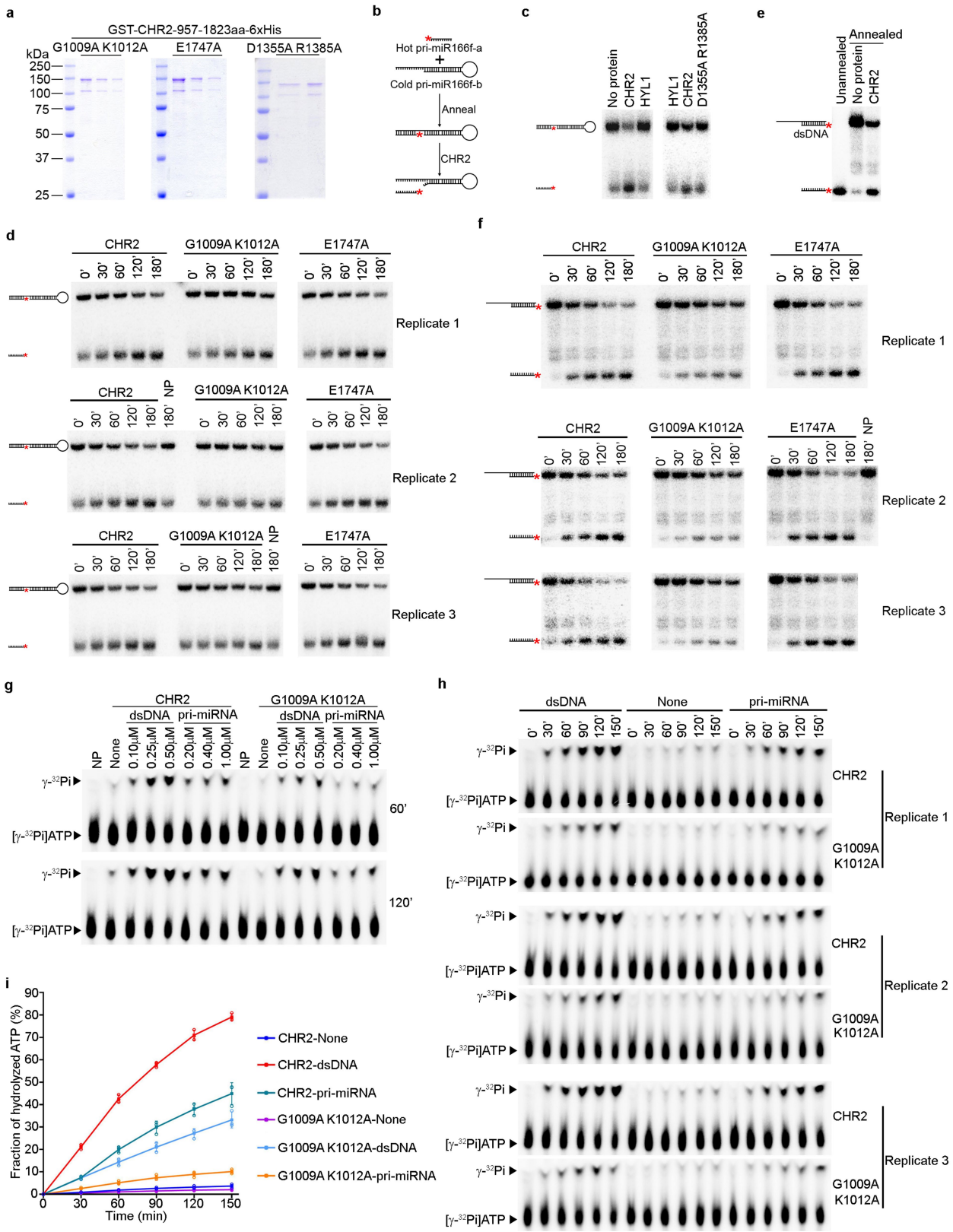


Extended Data Fig. 8 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

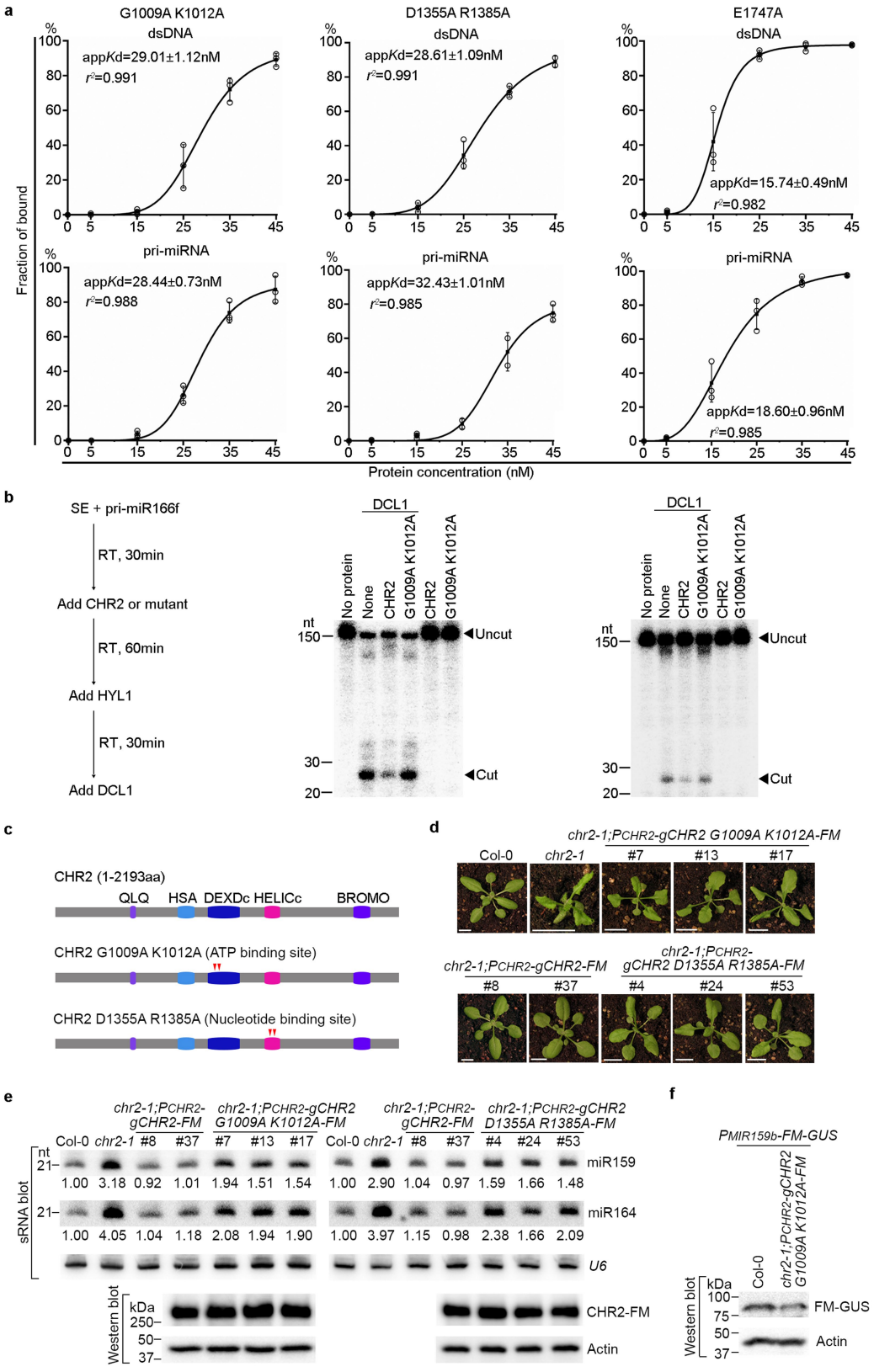
Extended Data Fig. 8 | CHR2 binds to nucleic acids in vitro and in vivo. **a**, EMSA shows CHR2 bound strongly to 5′ end ^{32}P-labelled pre-miRNA but weakly to ssRNA. Arrows indicate the mobility of protein−RNA complexes or labelled free RNA. **b**, **c**, EMSA shows the mobility pattern of the CHR2−pri-miRNA (**b**) and CHR2−pre-miRNA complexes (**c**). HYL1 and hSTING are positive and negative controls, respectively. Cold probes, used for the binding competition, are unlabelled pri-miRNA and pre-miRNAs, accordingly. **d**, **e**, EMSA shows that CHR2 bound to dsDNA but not ssDNA. Arrows indicate the mobility of protein–DNA complexes or free DNA. **f**, The binding affinity (apparent K_d) of CHR2-dsDNA was calculated from EMSA image quantification with s.d. from three replicates. Individual data points are shown. **g**, EMSA shows the mobility pattern of SE–pri-miRNA complex. **h**, EMSA shows the mobility pattern of HYL1-pri-miRNA complex. **i**, EMSA shows that HYL1 readily sequestered pri-miRNA from SE, CHR2 or CHR2−SE complexes. **j**, Western blot analysis of CHR2−FM protein immunoprecipitated from *hyl1-2 chr2-1;PCHR2-gCHR2-FM* plants using an anti-Myc antibody. Histone 3 serves as a control for input. **k**, RT−qPCR shows that none or faint signals were detected in the RNase-treated RIP samples. The result indicated that the nucleic acids recovered from the CHR2 immunoprecipitate shown in Fig. 4d were indeed RNAs. The relative signals of pri-miRNAs were normalized to the ones of the input initially and then to the control immunoprecipitate without RNase treatment, where the ratio was arbitrarily set to 1 with s.d. from three biological repeats. Individual data points are shown. ND, not detected.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | Effect of CHR2 mutations on remodelling/helicase and ATPase activities. a, SDS–PAGE of purified CHR2 variants. **b**, Schematic diagram of preparation of the ^{32}P-labelled and nicked pri-miRNA. **c**, Remodelling/helicase activity of CHR2 and CHR2(D1355A/R1385A) on nicked pri-miRNAs. Note: D1355A R1385A substitutions in the nucleotide binding domains compromised the remodelling activity of CHR2 on pri-miRNA. HYL1 served as a negative control. **d**, Time-course of remodelling/helicase activity of CHR2 and its variants on ^{32}P-labelled pri-miRNA. **e**, Remodelling/helicase activity of CHR2 on ^{32}P-labelled dsDNA. **f**, Time-course of remodelling/helicase activity of CHR2 and its variants on ^{32}P-labelled dsDNA. **g**, TLC assays showed that dsDNA and pri-miRNA stimulated ATPase activity of CHR2 and its ATPase mutant (G1009A/K1012A) in a dosage-dependent manner. **h**, TLC assays show the time-course of ATPase activity of CHR2 and its ATPase mutant (G1009A/K1012A) in the absence (none) or presence of dsDNA (167 bp in length) and pri-miRNA (150 nt in length with ~50 base-pairing). **i**, Quantification of ATP hydrolysis rates of CHR2 and its ATPase mutant (G1009A/K1012A) in the absence (none) or presence of dsDNA and pri-miRNA with s.d. from three replicates. Individual data points are shown.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

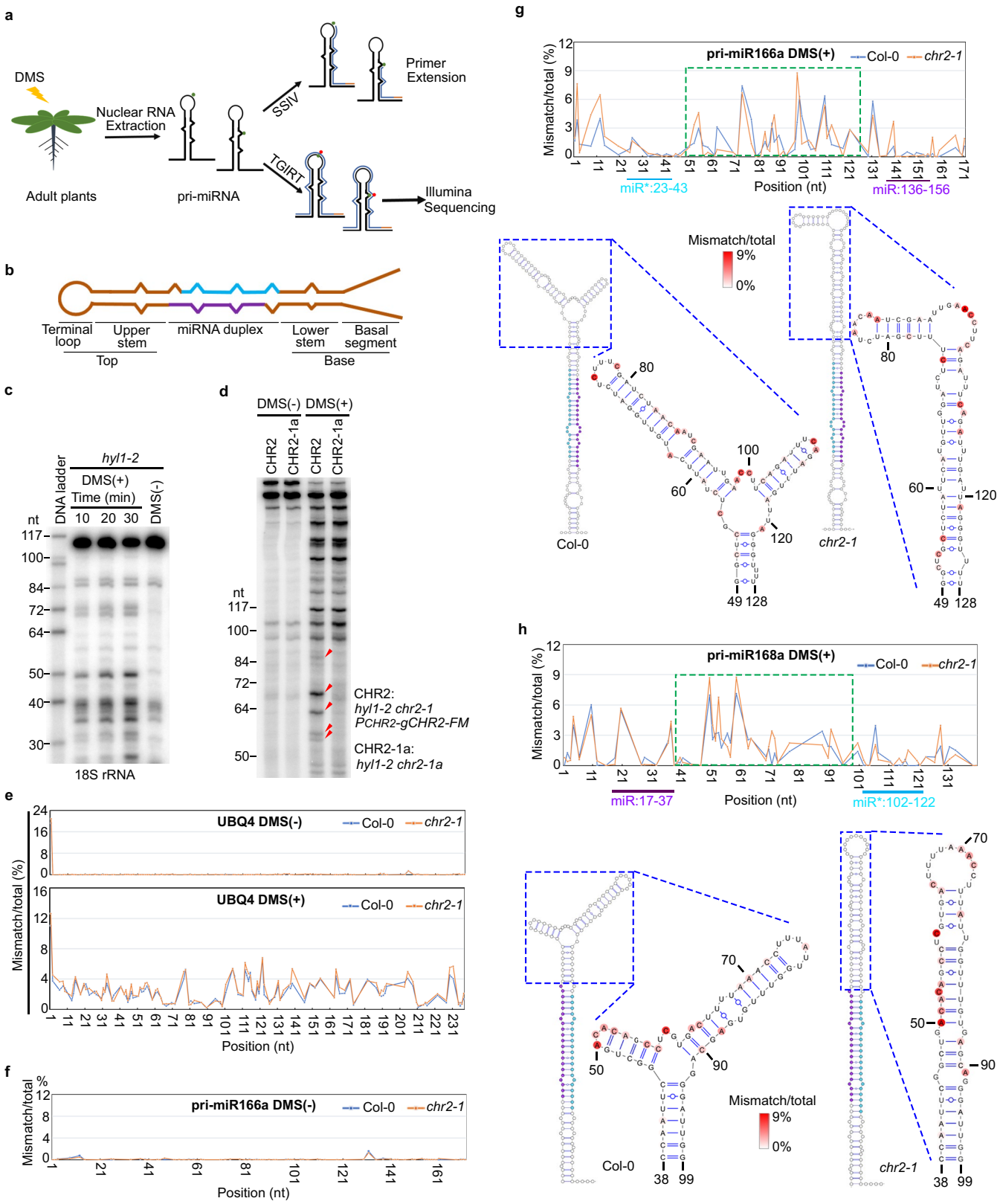


Extended Data Fig. 10 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 10 | Effect of CHR2 mutations on binding affinity to nucleic acids, Microprocessor cleavage efficiency of pri-miRNA, and rescue of morphological and miRNA accumulation defects in *chr2-1* plants. a, Binding curves of CHR2 variants with dsDNA and pri-miRNA. Apparent K_d values (appK_d) were calculated from EMSA image quantification with s.d. from three replicates. **b**, Microprocessor assays shows that the remodelling-compromised CHR2 mutant (G1009A/K1012A) failed to efficiently inhibit pri-miRNA processing. Left, schematic illustration of in vitro reconstitution orders with recombinant CHR2 or its variant, 5′ ^{32}P-labelled pri-miR166f transcript and Microprocessor components. RT, room temperature. Middle and right panels, autoradiography of the in vitro Microprocessor assays (two replicates). Black arrows indicate processed and unprocessed fragments of pri-miR166f. **c**, Diagram of point mutations in the CHR2 ATP and nucleotide binding domains. **d**, Remodelling-compromised CHR2 mutants partially rescued the morphological defect of *chr2-1* plants. Scale bar, 1 cm. **e**, Remodelling-compromised CHR2 mutants partially rescued *chr2-1* miRNA accumulation defects. sRNA blot analysis (top) and western blot assays of CHR2 and variants (bottom). **f**, Western blot analysis of GUS protein in *PMIR159b-FM-GUS* homozygotes ($n > 30$) in different backgrounds shows that CHR2(G1009A/K1012A) decreased *MIR159b* transcription.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

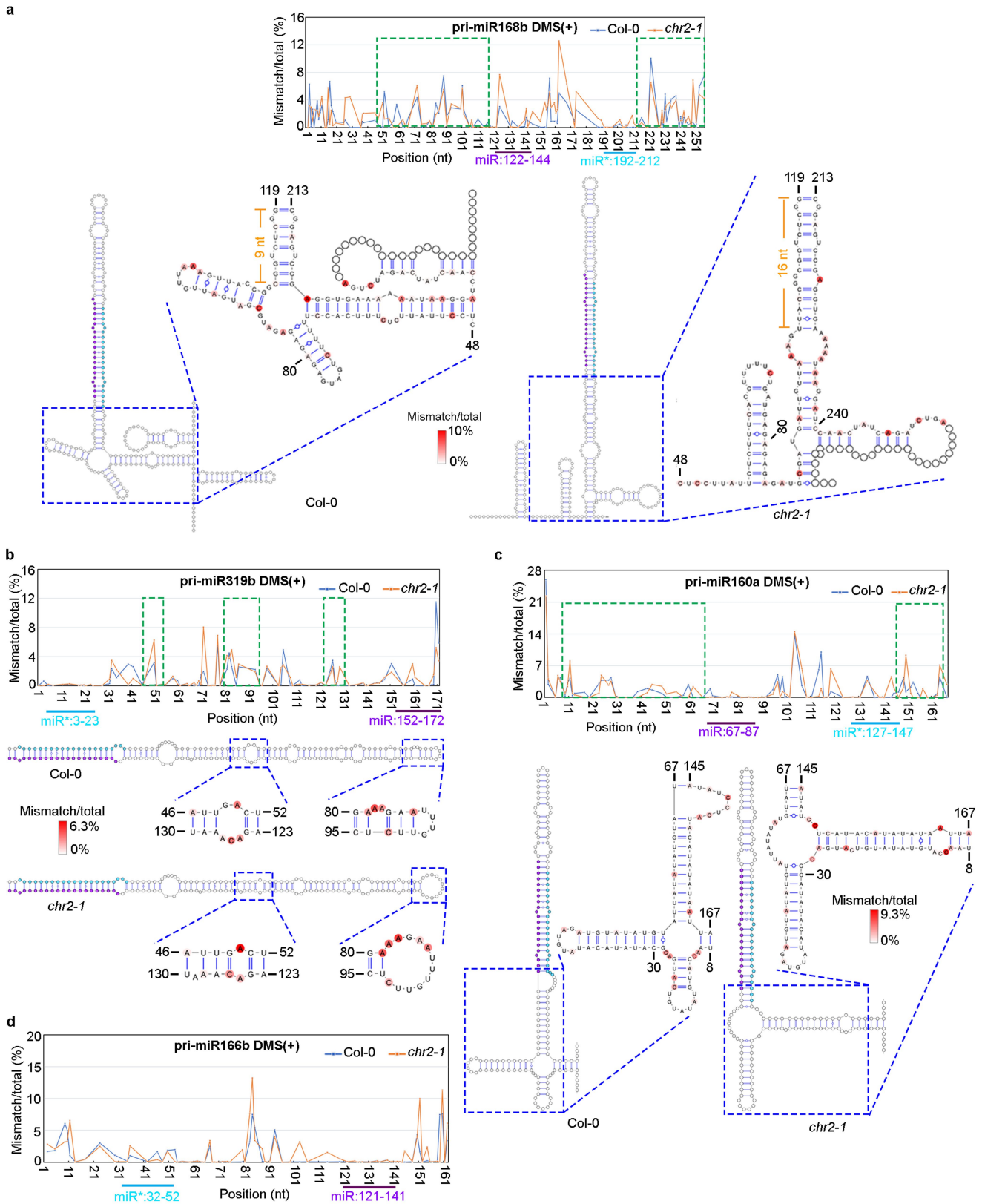


Extended Data Fig. 11 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 11 | CHR2 remodels the folding of pri-miRNAs or pri-miRNA complexes in vivo in various ways. a, Schematic diagram of two DMS-based methods for probing the secondary structures of RNA in vivo. SSIV, reverse transcriptase SuperScript IV (Thermo Fisher). **b**, Schematic structure of pri-miRNA. **c**, A primer-extension assay of DMS−cDNA truncations of 18S rRNA validated the reliability of the approach of the DMS−primer extension in the experiment. **d**, An additional replicate of primer-extension assay for DMS−cDNA truncations shows that CHR2 altered the folding of the miRNA/* duplex and upper stem region of pri-miR164b. Red arrows, nucleotides that had stronger DMS activities and thus more mispairings in pri-miR164b from *hyl1-2 chr2-1*;*PCHR2-gCHR2-FM* plants than those from *hyl1-2 chr2-1a*. **e**, DMS−MaPseq analysis shows that DMS mutation profiling of *UBQ4* transcripts was essentially identical in Col-0 and *chr2-1* plants. The DMS mutation frequencies (mismatch/total) of A and C residues are plotted along the nucleotide sequences probed. Background mutations in the samples that were not treated with DMS are shown (top). **f**, Background mutations in pri-miR166a in Col-0 and *chr2-1* samples that were not treated with DMS are shown. DMS−MaPseq results of other tested pri-miRNAs including pri-miR159a, pri-miR159b, pri-miR160a, pri-miR166b, pri-miR168a, pri-miR168b, and pri-miR319b in the untreated samples were largely identical to that of pri-miR166a, and are not shown here. **g**, **h**, DMS–MaPseq analysis of additional pri-miRNAs illustrates that CHR2 can alter nucleotide pairings and/or protein protection in terminal loops and upper stems of pri-miR166a (**g**) and pri-miR168a (**h**). Top, DMS mutation frequencies of A and C residues (mismatch/total) from targeted pri-miRNAs averaged from two biological replicates are plotted along pri-miRNA sequence. Position 1 corresponds to the 5′ end of the modelled regions. Bottom, overall and zoomed secondary structures of the targeted pri-miRNAs are predicted from DMS activities. Colour-coded nucleotides display different DMS activity on pri-miRNAs between Col and *chr2-1* plants. The miRNA and * strands are marked in purple and cyan, respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 12 | Additional evidence for the CHR2 function in remodelling the folding of pri-miRNAs or pri-miRNA complexes in vivo in various ways. a–d, DMS–MaPseq analysis of additional pri-miRNAs illustrates that CHR2 can alter nucleotide pairings and/or protein protection in the lower stem of pri-miR168b (**a**), the terminal loop and upper stem of pri-miR319b (**b**), and variable parts of pri-miR160a (**c**) and pri-miR166b (**d**). Top, DMS mutation frequencies of A and C residues (mismatch/total) from targeted pri-miRNAs averaged from two biological replicates are plotted along pri-miRNA sequence. Position 1 corresponds to the 5′ end of the modelled regions. Bottom, the overall and zoomed secondary structures of the targeted pri-miRNAs are predicted from DMS activities. Colour-coded nucleotides display different DMS activity on pri-miRNAs between Col and *chr2-1* plants. The miRNA and * strands are marked in purple and cyan, respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0133-z

Pulsar emission amplified and resolved by plasma lensing in an eclipsing binary

Robert Main[1,2,3]*, I-Sheng Yang[3,4], Victor Chan[1], Dongzi Li[3,5], Fang Xi Lin[3,5], Nikhil Mahajan[1], Ue-Li Pen[3,6,2,4], Keith Vanderlinde[1,2] & Marten H. van Kerkwijk[1]

Radio pulsars scintillate because their emission travels through the ionized interstellar medium along multiple paths, which interfere with each other. It has long been realized that, independent of their nature, the regions responsible for the scintillation could be used as 'interstellar lenses' to localize pulsar emission regions[1,2]. Most such lenses, however, resolve emission components only marginally, limiting results to statistical inferences and detections of small positional shifts[3–5]. As lenses situated close to their source offer better resolution, it should be easier to resolve emission regions of pulsars located in high-density environments such as supernova remnants[6] or binaries in which the pulsar's companion has an ionized outflow. Here we report observations of extreme plasma lensing in the 'black widow' pulsar, B1957+20, near the phase in its 9.2-hour orbit at which its emission is eclipsed by its companion's outflow[7–9]. During the lensing events, the observed radio flux is enhanced by factors of up to 70–80 at specific frequencies. The strongest events clearly resolve the emission regions: they affect the narrow main pulse and parts of the wider interpulse differently. We show that the events arise naturally from density fluctuations in the outer regions of the outflow, and we infer a resolution of our lenses that is comparable to the pulsar's radius, about 10 kilometres. Furthermore, the distinct frequency structures imparted by the lensing are reminiscent of what is observed for the repeating fast radio burst FRB 121102, providing observational support for the idea that this source is observed through, and thus at times strongly magnified by, plasma lenses[10].

On 2014 June 13–16, we took 9.5 h of data of PSR B1957+20 with the 305-m William E. Gordon Telescope at the Arecibo observatory, at observing frequency of 311.25–359.25 MHz (see Methods). These data were previously searched for giant pulses[11]—sporadically occurring, extremely bright pulses, which are much shorter than regular pulses (around 1 μs compared with tens of microseconds). Although we found many giant pulses at all orbital phases, we noticed that the incidence rate of bright pulses was much higher leading up to and following the radio eclipse.

As can be seen in Fig. 1, most of the pulses near eclipse do not look like giant pulses but rather like brighter regular pulses—they are bright over a large fraction of the pulse profile, and most tellingly, occur in groups spanning several 1.6-ms pulse rotations, suggesting that the underlying events last for a time of the order of 10 ms. Their properties seem similar to the hitherto mysterious bright pulses associated with the eclipse of PSR J1748 − 2446A[12], suggesting a shared physical mechanism.

High-magnification events are often chromatic, as can be seen from the colours in Fig. 1 (which reflect 16-MHz sub-bands) and is borne out more clearly by the spectra shown in Fig. 2: some show frequency widths comparable to our 48-MHz band, peaking at low or high frequency, whereas others show strong frequency evolution, in some cases tracing out a slope in frequency–time space, in others a double-peaked profile.

For many high-magnification events, the magnification is not uniform across the pulse profile. The bottom panel of Fig. 1 shows this strikingly: at 0.2 s, the main pulse is greatly magnified over five pulses, whereas the entire interpulse is barely affected. In addition, the components of the broad interpulse are often magnified differently from each other, as can be seen most readily in the magnified events in Fig. 3b, c. Thus, the events resolve the pulsar's various emission regions.

The most strongly magnified events occur within three specific time spans, each lasting for about 5 min, around orbital phase $\phi = 0.20$, $\phi = 0.30$ and $\phi = 0.32$ (see Fig. 1); here, $\phi = 0.25$ corresponds to superior conjunction of the pulsar in its 9.2-h orbit[7], and the duration of the eclipses at 350 MHz is about 40 to 60 min (refs. [7–9]), or $\Delta\phi \approx 0.1$ (a sketch of the system geometry is shown in Extended Data Fig. 1). In each time span, we observe only small overall frequency-dependent time delays, of the order of 10 μs, which indicate modest increases in electron column density, with dispersion measures of the order of 10^{-4} pc cm^{-3} (see Fig. 1 as well as Extended Data Fig. 2). Intriguingly, at times when the delays are longer and the dispersion measures thus higher—immediately before and after eclipse, and in between the two post-eclipse periods of strong lensing, at $\phi \approx 0.31$—the magnifications are less marked, only up to a factor of a few, and correlated over much longer timescales, of the order of 100 ms. After the main lensing periods, up to $\phi \approx 0.36$, flux variations remain correlated, suggesting that even weaker events are still present (examples of lensing in the aforementioned regions are shown in Extended Data Fig. 3).

To determine whether the magnification events could be due to lensing by inhomogeneities in the companion's outflow, we first measured the excess delay due to dispersion at a time resolution of 2 s, the shortest at which we can measure it reliably (see Methods). We find that during the periods of strong magnification events, the delay fluctuates by about 1 μs on this timescale. Assuming that the relative velocity between the pulsar and the companion's outflow is roughly the orbital velocity, 360 km s^{-1} (ref. [13]), the 2-s timescale corresponds to a spatial scale $\Delta x \approx 720$ km. For this scale, the expected geometric delay is $\Delta x^2/2ac \approx 0.5$ μs (where $a = 6.4$ light-seconds is the orbital separation and c the speed of light). As this is comparable to the observed dispersive delays, lensing is expected.

To model the magnified pulses, we treat the signal using a standard wave-optics formalism. The electric field received by an observer is the sum of the electric field of the source across the lens plane, with the phase at every point determined by both geometric and dispersive delays. When a large area on the plane has (nearly) stationary phase, the electric field combines coherently, leading to a strongly magnified image, with the magnification μ proportional to the area squared. Because dispersive and geometric delays scale differently with frequency ν, a plasma lens cannot be in focus over all frequencies, but will impart a characteristic frequency width. This width will be smaller for larger magnification, with the precise scaling depending on the extent to which the lens is elliptic: $\Delta\nu/\nu \sim 1/\mu$ for a very elongated, effectively linear lens, and $\Delta\nu/\nu \sim 1/\sqrt{\mu}$ for a (roughly) circular one. In strong lensing, one generically expects multiple images to contribute, and thus

[1]Department of Astronomy and Astrophysics, University of Toronto, Toronto, ON, Canada. [2]Dunlap Institute for Astronomy and Astrophysics, University of Toronto, Toronto, ON, Canada. [3]Canadian Institute for Theoretical Astrophysics, University of Toronto, Toronto, ON, Canada. [4]Perimeter Institute for Theoretical Physics, Waterloo, ON, Canada. [5]Department of Physics, University of Toronto, Toronto, ON, Canada. [6]Canadian Institute for Advanced Research, Toronto, ON, Canada. *e-mail: main@astro.utoronto.ca

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

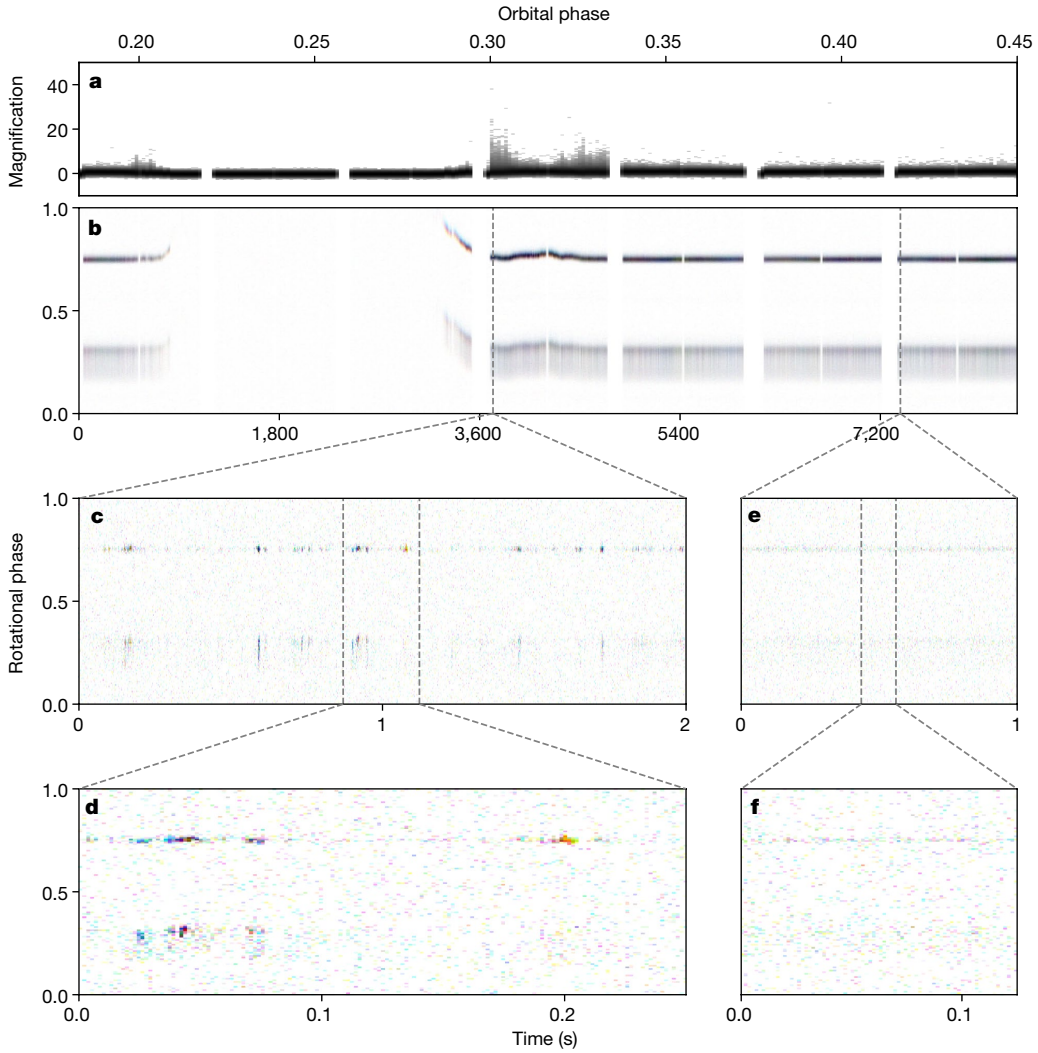


Fig. 1 | Strong lensing of the pulsar. a, Magnification distributions for the main pulse, in 10-s time bins, showing periods of lensing near ingress and egress of the eclipse of the pulsar by its companion's outflow. The greyscale is logarithmic, to help individual bright pulses to stand out. **b**, Pulse profiles as a function of time, averaged over 10 s and in 128 phase bins. Cyan, yellow and magenta represent contiguous 16-MHz sub-bands, from low to high frequency. Near eclipse, plasma dispersion causes frequency-dependent delays. Gaps in the data correspond to calibrator scans. **c**, **d**, Enlargements in which each time bin corresponds to an individual 1.6-ms pulse. One sees extreme chromatic lensing in which pulses are magnified by an order of magnitude over tens of milliseconds; the brightest event is magnified by a factor of about 40 across our highest-frequency sub-band. In some events, the main pulse and interpulse are clearly affected differently, indicating that different emission regions corresponding to these are resolved by the lensing structures. **e**, **f**, Enlargements for a quiescent period for comparison.

caustics to form, which can lead to a slope in time–frequency space, double-peaked spectra and other interference effects (see Methods). All of these are consistent with the behaviour seen in the measured spectra of the magnified pulses in Fig. 2.

For any lens, different pulse components will be magnified differently if they arise from regions that have projected separations larger than the lens's resolution. We can make a quantitative estimate of the lens resolution by using the magnifications of the events. Continuing to assume the lens is roughly co-located with the companion, that is, at the orbital separation a, the resolution of the lens for given magnification would be $\approx 1.9R_1/\mu^{1/2}$ for a linear lens, or $\approx 1.9R_1/\mu^{1/4}$ for a circular one, where $R_1 \equiv \sqrt{\lambda a/\pi}$, or 23 km at our observing wavelength $\lambda = c/\nu \approx 90$ cm (see Methods for a derivation). A peak magnification of, for instance, $\mu = 50$ then corresponds to a physical resolution of about 6 km for a linear lens or about 17 km for a circular one.

The inferred resolution is comparable to the radius (approximately 10 km) of the neutron star, and substantially smaller than the light-cylinder radius, $R_{\rm LC} \equiv cP/2\pi = 76$ km, where the velocity of co-rotating magnetic fields approaches the speed of light, bounding the magnetosphere in which pulsar emission is thought to originate[14]. Thus, the lensing offers the opportunity to map the emission geometry.

Qualitatively, because the main pulse and interpulse, as well as parts of the wide interpulse beam, are sometimes magnified very differently, the inferred resolution is a lower limit both to the projected separation between the main pulse and the interpulse, and to the size of the interpulse. Perhaps not unexpectedly, the spatial separations do not seem to map directly to rotational phase: we see similar differences between the main pulse and the interpulse, which are separated by half a rotation, and between parts of the interpulse, which are separated by only about 0.1 rotation. This may indicate that the interpulse consists of multiple components, which are not located close together in space.

Combining the lens resolution with the timescale of the lensing events allows us to constrain the projected relative velocity between the lens and the pulsar. A priori, one might expect the outflow velocity to be slow, in which case the relative velocity would just be the orbital motion of about 360 km s^{-1}. Given that strong magnification events typically last about 10 ms, this would then imply a resolution of approximately

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

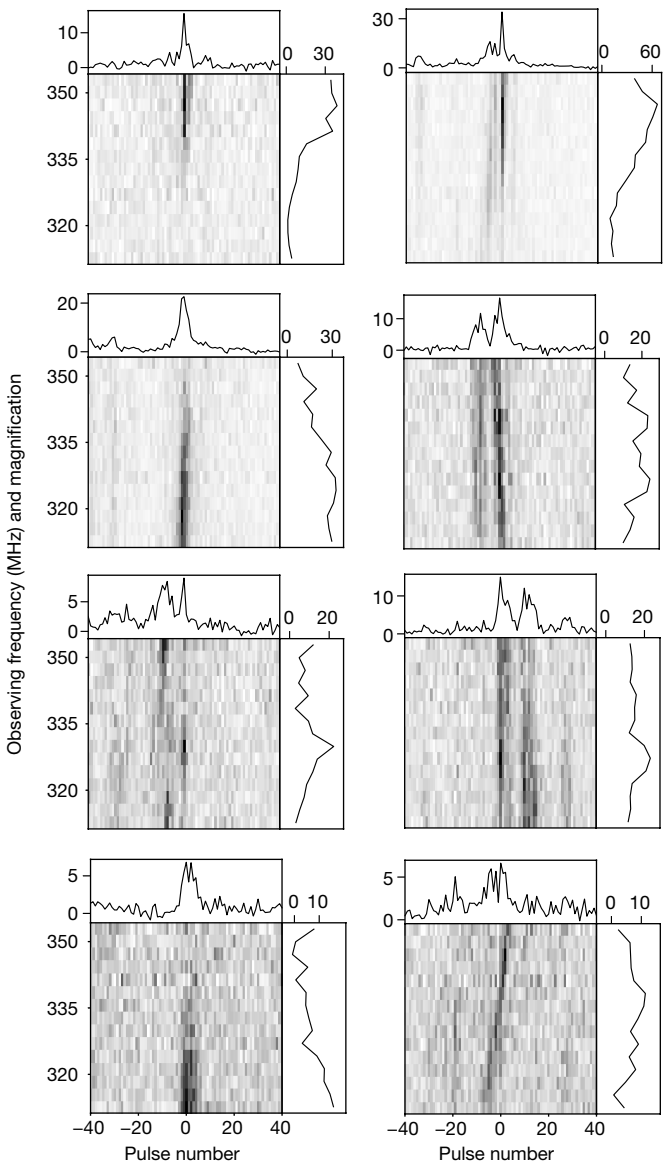


Fig. 2 | A range of spectral behaviour in lensing events. Power spectra, binned to 3 MHz, of the main pulse in consecutive 1.6-ms rotations, for selected lensing events from eclipse egress. The top panels show the average magnification in each main pulse, and the side panels the magnification spectra of the brightest pulse (that is, that of pulse number 0).

4 km, not dissimilar from what we find above, thus suggesting that the assumption of a slow outflow velocity is reasonable. By combining different constraints between the duration and frequency widths of lensing events, we can set a quantitative limit to the relative velocity (see Methods for a derivation), of

$$v > 0.5 \frac{R_1}{\Delta t_{\mathrm{HWHM}}} \left(\frac{\Delta \nu}{\nu} \right)^{1/2}_{\mathrm{HWHM}} > 360 \ \mathrm{kms}^{-1} \qquad (1)$$

where for the numerical value we use the fact that the tightest constraints come from the shortest, least chromatic events, which have frequency widths $(\Delta\nu/\nu)_{\mathrm{HWHM}} \approx 0.1$, and durations $\Delta t_{\mathrm{HWHM}} \approx 10$ms (measured as half-width at half-maximum, HWHM; see Fig. 2). This approximate limit equals the relative orbital velocity, which implies that the outflow velocity can be (but does not have to be) small in the rest frame of the companion.

Our results offer many avenues of further research. Ideally, one would use the lensing to map the pulsar magnetosphere. This requires better constraints on the lenses, for example from observations over several eclipses at a range of frequencies. Further observations may also shed light on the eclipse mechanism. For example, if it is cyclo-synchrotron absorption, large magnetic fields, of about 20 G, are required[15], which would impart measurable polarization dependence of the lensing events. Furthermore, combining density and velocity into

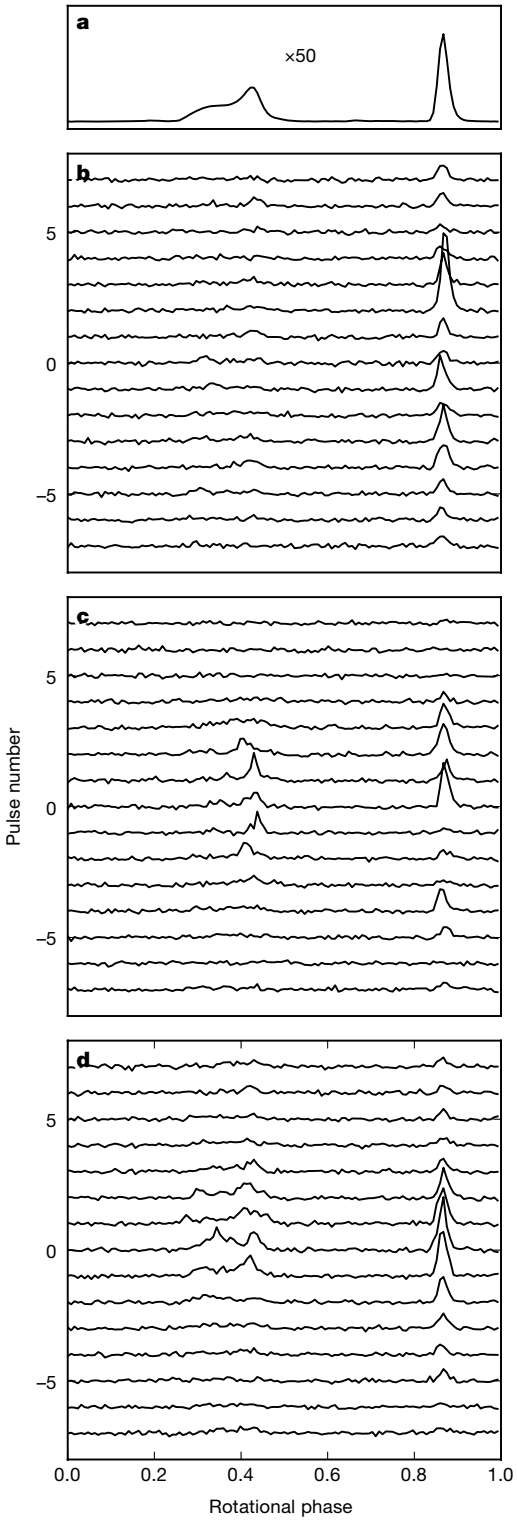


Fig. 3 | Profiles of lensing events. a, Frequency-averaged pulse profile of a 9-min quiescent region, scaled up by a factor of 50 for comparison purposes; **b, c, d**, frequency-averaged pulse profiles surrounding bright events (using 128 phase bins). One sees that the events resolve the magnetosphere: the main pulse and interpulse are affected differently, as are parts within the interpulse.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

a mass-loss rate of the outflow, one can infer the lifetime, and thus the final fate, of these systems.

Finally, our observations establish that radio pulses can be strongly amplified by lensing in local ionized material. This adds support for the proposal that fast radio bursts (FRBs) can be lensed by host galaxy plasma, leading to variable magnification, narrow frequency structures and clustered arrival times of highly amplified (and thus observable) events[10]. Evidence for high-density environments includes that FRB 110523 is scattered within its host galaxy[16], and that FRB 121102 has been localized to a star-forming region[17,18], in an extreme and dynamic magneto-ionic environment[19]. Indeed, the latter's repeating bursts have spectra[20–22] remarkably similar to those shown in Fig. 2 (for example, compare with Fig. 2 of ref. [20]), and, like those bursts, the brightest pulses in PSR B1957+20 are highly clustered in time.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0133-z.

Received: 7 December 2017; Accepted: 5 March 2018;
Published online 23 May 2018.

1. Lovelace, R. V. E. *Theory and Analysis of Interplanetary Scintillations*. PhD thesis, Cornell Univ. (1970).
2. Backer, D. C. Interstellar scattering of pulsar radiation. I: Scintillation. *Astron. Astrophys.* **43**, 395–404 (1975).
3. Gwinn, C. R. et al. Size of the Vela pulsar's emission region at 18 cm wavelength. *Astrophys. J.* **758**, 7 (2012).
4. Johnson, M. D., Gwinn, C. R. & Demorest, P. Constraining the Vela pulsar's radio emission region using Nyquist-limited scintillation statistics. *Astrophys. J.* **758**, 8 (2012).
5. Pen, U.-L., Macquart, J.-P., Deller, A. T. & Brisken, W. 50 picoarcsec astrometry of pulsar emission. *Mon. Not. R. Astron. Soc.* **440**, L36–L40 (2014).
6. Main, R. *et al*. Mapping the emission location of the Crab pulsar's giant pulses. Preprint at https://arxiv.org/abs/1709.09179 (2017).
7. Fruchter, A. S., Stinebring, D. R. & Taylor, J. H. A millisecond pulsar in an eclipsing binary. *Nature* **333**, 237–239 (1988).
8. Fruchter, A. S. et al. The eclipsing millisecond pulsar PSR 1957 + 20. *Astrophys. J.* **351**, 642–650 (1990).
9. Ryba, M. F. & Taylor, J. H. High-precision timing of millisecond pulsars. II: Astrometry, orbital evolution, and eclipses of PSR 1957 + 20. *Astrophys. J.* **380**, 557–563 (1991).
10. Cordes, J. M. et al. Lensing of fast radio bursts by plasma structures in host galaxies. *Astrophys. J.* **842**, 35 (2017).
11. Main, R., van Kerkwijk, M., Pen, U.-L., Mahajan, N. & Vanderlinde, K. Descattering of giant pulses in PSR B1957+20. *Astrophys. J. Lett.* **840**, L15 (2017).
12. Bilous, A. V., Ransom, S. M. & Nice, D. J. in *American Institute of Physics Conf. Ser.* Vol. 1357 (eds Burgay, M. et al.) 140–141 (2011).
13. Van Kerkwijk, M. H., Breton, R. P. & Kulkarni, S. R. Evidence for a massive neutron star from a radial-velocity study of the companion to the black-widow pulsar PSR B1957+20. *Astrophys. J.* **728**, 95 (2011).
14. Ginzburg, V. L. & Zhelezniakov, V. V. On the pulsar emission mechanisms. *Annu. Rev. Astron. Astrophys.* **13**, 511–535 (1975).
15. Thompson, C., Blandford, R. D., Evans, C. R. & Phinney, E. S. Physical processes in eclipsing pulsars: eclipse mechanisms and diagnostics. *Astrophys. J.* **422**, 304–335 (1994).
16. Masui, K. et al. Dense magnetized plasma associated with a fast radio burst. *Nature* **528**, 523–525 (2015).
17. Tendulkar, S. P. et al. The host galaxy and redshift of the repeating fast radio burst FRB 121102. *Astrophys. J. Lett.* **834**, L7 (2017).
18. Bassa, C. G. et al. FRB 121102 is coincident with a star-forming region in its host galaxy. *Astrophys. J. Lett.* **843**, L8 (2017).
19. Michilli, D. et al. An extreme magneto-ionic environment associated with the fast radio burst source FRB 121102. *Nature* **553**, 182–185 (2018).
20. Spitler, L. G. et al. A repeating fast radio burst. *Nature* **531**, 202–205 (2016).
21. Scholz, P. et al. The repeating fast radio burst FRB 121102: multi-wavelength observations and additional bursts. *Astrophys. J.* **833**, 177 (2016).
22. Law, C. J. et al. A multi-telescope campaign on FRB 121102: implications for the FRB population. *Astrophys. J.* **850**, 76 (2017).

Acknowledgements We thank A. Bilous for sharing and discussing her results which motivated this analysis. We thank the rest of our scintillometry group for discussions and support throughout. R.M., M.H.v.K., U.-L.P. and K.V. are funded through NSERC. I-S.Y was supported by a SOSCIP Consortium Postdoctoral Fellowship.

Reviewer information Nature thanks S. Ransom and J. Hessels for their contribution to the peer review of this work.

Author contributions R.M. discovered the magnified pulses; wrote the majority of the manuscript; and created the figures. I-S.Y. developed and wrote the sections on wave-optics formalism in the Methods; M.H.v.K and U.-L.P. guided the analysis and interpretation of the results; M.H.v.K. also helped to write the manuscript and influenced the content and presentation of the figures; F.X.L., V.C. and K.V. quantified the excess dispersion measure associated with the eclipse. N.M. wrote the code used to produce the coherently de-dispersed single pulse profiles used in this analysis, and derived an improved orbital ephemeris relative to which time delays are measured. D.L. worked on criteria to systematically separate giant pulses from pulses magnified by plasma lensing. All authors contributed to the interpretation of results.

Competing interests The authors declare no competing financial interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0133-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to R.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Calculating magnifications. The data were taken in four 2.4-h sessions on 2014 June 13–16 and were recorded as part of a European VLBI network programme (GP 052). The data span 311.25–359.25 MHz, in three contiguous 16-MHz sub-bands, and we read, de-dispersed (with a dispersion measure of 29.1162 pc cm^{-3}) and reduced the data as described in ref. [11]. As one extra step, we accounted for the wander in orbital period of PSR B1957+20 by adjusting the time of ascending node in the ephemeris, such that it minimized the scatter of the arrival times of giant pulses associated with the main pulse across all four days of observation.

We define the magnification of a pulse as the ratio of its flux to the mean flux in a quiescent region far from eclipse. Specifically, we construct an intensity profile of each pulse using 128 phase gates, subtract the mean off-pulse flux in each 16-MHz sub-band separately, measure the flux in an eight-gate (approximately 100 μs) window around the peak location (which we find from folded profiles, averaged in 2-s bins to obtain sufficient signal-to-noise), and divide by the average flux in the same pulse window measured in a 9-min section far from eclipse.

To construct the spectra of lensing events, we start by binning power spectra in 3-MHz bins. We correct these approximately for the bandpass and the effects of interstellar scintillation (which still has a small amount of power on 3-MHz scales) by dividing them by the average spectra in the 15 s before and after (excluding the lensing event itself). This time span is chosen to be safely less than the timescale of about 84 s on which the interstellar scintillation pattern varies[11], ensuring the dynamic spectrum is stable on these scales. With 30 s of data, it is also very well measured: each 3-MHz channel has signal-to-noise ratio $S/N \approx 150$.

Evidence that strong plasma lensing must occur. To obtain the properties of the lensed images, and in particular to determine whether we are in the 'strong' or 'weak' lensing regime—that is, whether or not multiple images are formed—we use the basic principles of wave optics, considering path integrals of the electric field over a thin lens[23]. In our case, the phase of an electromagnetic wave going through different paths has contributions from both geometric and dispersive time delays,

$$\phi(x,y) = \phi_{\rm GM}(x,y) + \phi_{\rm DM}(x,y) \tag{2}$$

where x and y are in the lens plane. For a geometrically thin lens, that is, with thickness along the line of sight much smaller than the separation a, the geometric contribution to the phase can be written as

$$\phi_{\rm GM} = \frac{1}{2}\left(\frac{x^2+y^2}{a^2}\right)\left(\frac{a}{\lambda}\right) = \frac{x^2+y^2}{2R_{\rm Fr}^2} \tag{3}$$

where $R_{\rm Fr}$ is the Fresnel scale,

$$R_{\rm Fr} \equiv \sqrt{\lambda a} \approx 40\ {\rm km} \tag{4}$$

with λ the wavelength of the radiation. For the numerical value, we used $\lambda = c/\nu = 90$ cm (where ν is the observing frequency) and assumed that the lensing material was associated with the companion and thus at the orbital separation of $a \approx 6.4$ light-seconds. (Note that as we are using wavelengths and frequencies of full cycles, our phases are in cycles too, not radians.)

The dispersive contribution arises from the signal propagating through the lens's extra dispersion measure (electron column density) $\Delta{\rm DM} = \int n_{\rm e}{\rm d}z$ (with $n_{\rm e}$ the electron number density),

$$\phi_{\rm DM}(x,y) = -\frac{k_{\rm DM}}{\nu}\Delta{\rm DM}(x,y) \tag{5}$$

where $k_{\rm DM} = e^2/2\pi m_{\rm e}c = 4{,}148.808$ s pc^{-1} cm^3 MHz^2 (ref. [24]). The minus sign arises because in a plasma the phase velocity is greater than the speed of light.

Integrating over different paths, one effectively select regions where the electric fields are coherent; these lead to the final images. For instance, if the extra electrons are distributed uniformly (in the x–y plane after the z integral), then $\phi_{\rm DM}$ is approximately constant and the total phase has a stationary point around $x = y = 0$, that is, around the line of sight. Furthermore, all paths that are less than the Fresnel scale away from this central path have similar phase, and thus one recovers the general result that the area of the lens that contributes scales with $R_{\rm Fr}^2$.

For the lensing to be strong, a minimum requirement is that changes in the geometric and dispersive phase are of similar magnitude. Because dispersion also leads to overall time delays, differences in pulse arrival time give a measure of inhomogeneities in the lensing material. We measure pulse arrival times using the usual procedure of fitting pulse profiles to a high signal-to-noise template (from a quiescent period). We fit profiles in the three bands separately and convert the weighted average to ΔDM using standard equations. We find that the inferred ΔDM shows both the expected large-scale variations[7] and considerable variability down to the shortest timescales (2 s) at which we can measure it reliably (see Extended Data Fig. 2). During the periods in which the strongly magnified events occur, we find intrinsic variability corresponding to variations in delay of $\Delta t_{\rm DM} \approx 1$ μs, which suggests that the inhomogeneities correspond to differences in the dispersion phase of $\Delta\phi_{\rm DM} = \nu\Delta t_{\rm DM} \approx 300$ cycles at 2-s timescales.

To compare this to the geometric phase, we need to translate the timescale of 2 s to a length scale. Assuming the relative motion between the pulsar and the lensing material is dominated by the orbital motion, of 360 km s^{-1}, the spatial scale is 720 km. Using equation (3), this corresponds to $\Delta\phi_{\rm GM}$ (720 km) ≈ 200 cycles.

The fact that $\Delta\phi_{\rm DM}$ is comparable to and slightly larger than $\Delta\phi_{\rm GM}$ at the distance scale of 720 km guarantees that somewhere around this scale, slopes in $\phi_{\rm DM}$ and $\phi_{\rm GM}$ will sometimes cancel. It also implies that multiple stationary points with $\nabla\phi = 0$ are likely. Because the phases result from a continuous function of x, y and ν, the stationary points must emerge or annihilate in pairs, leading to so-called caustics, where one has not just $\nabla\phi = 0$ but also $\det(\partial_i\partial_j\phi) = 0$ (ref. [25]).

Given the above, strong lensing is a natural cause of the strongly magnified events that we observe.

A perfect lens model. To calculate properties of the lensed images, we would need to perform the path integral. This is not possible, as it requires the value of $\phi_{\rm DM}$ of the full two-dimensional lens plane at a resolution better than the Fresnel scale, whereas our time-delay measurements are limited to the one-dimensional trajectory of the pulsar, and to an order of magnitude larger scales.

Hence, for our analysis we use a simplified model, amenable to wave-optics analysis, in which each strong lensing event is associated with an elliptical perfect lens, that is, there is a single focal point to which all paths within the lens contribute perfectly coherently, and all paths outside the lens do not contribute at all. As shown in Extended Data Fig. 4, our model is parametrized by the location of the focal point relative to the centre of the lens, (X, Y), the semi-major and semi-minor axes of the lens, ($\Delta x_{\rm lens}$, $\Delta y_{\rm lens}$) and the orientation relative to the projected pulsar trajectory.

In the special case of a centred circular lens with radius R, that is, $\Delta x_{\rm lens} = \Delta y_{\rm lens} = R$ and $X = Y = 0$, the set-up would produce an Airy disk. By comparing this to the actual path-integral of an unlensed image (that is, with $\phi_{\rm DM} = 0$), we can identify the unlensed case with a circular lens with radius $R_{\rm I} \equiv R_{\rm Fr}/\sqrt{\pi}$.

Integrating the electric field over an elliptical lens, the magnification (defined as the increase in intensity $\mu = I/\langle I\rangle$, where $I = E^2$), will be proportional to the area squared:

$$\mu = \left(\frac{\Delta x_{\rm lens}}{R_{\rm I}}\right)^2\left(\frac{\Delta y_{\rm lens}}{R_{\rm I}}\right)^2 \tag{6}$$

Before continuing, we note that at first glance it might seem puzzling for the magnification to be proportional to image area squared, as that seems to violate energy (flux) conservation. However, the image is also more highly beamed; what we calculated is the magnification in the centre of the beam. Viewed from an angle, a linear phase term is induced across the originally coherent region. Thus, when the region is larger, it is easier to become incoherent. The solid angle scales as $\Delta\Omega \propto \Delta x_{\rm lens}^{-1}\Delta y_{\rm lens}^{-1}$, and thus total energy (flux) is indeed proportional to the area of the lens.

Differential magnification and chromaticity. We now proceed to derive the physical resolution of the lens. The above magnification is reached only when the source is exactly at the focal point. When the emission region is some distance away from the focal point, paths from it to different parts of elliptical region will have extra phase differences. For instance, when a source is separated from the focal point in the x direction by $x_{\rm source} \equiv (X - x_{\rm s})$ (where $x_{\rm s}$ is measured relative to the centre of the lens), there is a phase difference across the lens of

$$\Delta\phi = \frac{\Delta x_{\rm lens} x_{\rm source}}{R_{\rm I}^2} \tag{7}$$

When this phase difference reaches order 1, the image from the new source location will no longer be magnified. Defining the resolution as the offset for which total cancellation happens, we find from an explicit integral over the elliptical lens[1]

$$x_{\rm res} \approx \frac{1.9R_{\rm I}^2}{\Delta x_{\rm lens}} \tag{8}$$

$$y_{\rm res} \approx \frac{1.9R_{\rm I}^2}{\Delta y_{\rm lens}} \tag{9}$$

Hence, total cancellation occurs on an elliptical beam with semi-minor and semi-major axes $x_{\rm res}$ and $y_{\rm res}$ that scale inversely to the semi-major and semi-minor axes of the lens.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

We can write the above in terms of the magnification for a specific lens model. We will consider two extremes, the first a very anisotropic lens, effectively one-dimensional, for which $\Delta y_{\rm lens} = R_{\rm I}$. For this case,

$$x_{\rm res} \approx \frac{1.9R_{\rm I}}{\mu^{1/2}} \tag{10}$$

In the opposite direction, we consider a circular lens, with $\Delta x_{\rm lens} = \Delta y_{\rm lens}$. For this case,

$$x_{\rm res} = y_{\rm res} \approx \frac{1.9R_{\rm I}}{\mu^{1/4}} \tag{11}$$

Here, dividing by the distance a and writing in terms of diameter $D = 2\Delta x_{\rm lens}$, one recovers the usual relation for the angular resolution of a circular lens, $1.22\lambda/D$.

A similar result can be derived in frequency space. Given the different frequency dependencies of the geometric and dispersive phases, a lens cannot be fully in focus across all frequencies, but will have some characteristic frequency width. To show this, we first assume that the focal point is co-located with the centre of the elliptical lens, that is, $(X, Y) = (0,0)$, and that perfect coherence occurs at some frequency $\nu_{\rm c}$, that is, that at $\nu_{\rm c}$ one has

$$\phi_{\rm GM}(x, y, \nu_{\rm c}) = -\phi_{\rm DM}(x, y, \nu_{\rm c}) = \frac{\nu_{\rm c}(x^2 + y^2)}{2ac} \tag{12}$$

The total phase at a nearby frequency is then given by

$$\phi(x, \nu_{\rm c} + \Delta\nu) = \left(\frac{\nu_{\rm c} + \Delta\nu}{\nu_{\rm c}} - \frac{\nu_{\rm c}}{\nu_{\rm c} + \Delta\nu}\right)\frac{\nu_{\rm c}(x^2 + y^2)}{2ac} \tag{13}$$

the dominant term of which is linear in $\Delta\nu$. If this term is of order 1 within the elliptical region, the image will no longer be magnified. Perfect cancellation does not happen in this case, but one can define characteristic width. Using half-width at half-maximum, one finds

$$\left(\frac{\Delta\nu}{\nu_{\rm c}}\right)_{\rm HWHM} = \frac{2.9R_{\rm I}^2}{\sqrt{3\Delta x_{\rm lens}^4 - 2\Delta x_{\rm lens}^2\Delta y_{\rm lens}^2 + 3\Delta y_{\rm lens}^4}} \tag{14}$$

In terms of the magnification, we find for the special case of an effectively one-dimensional lens

$$\left(\frac{\Delta\nu}{\nu_{\rm c}}\right)_{\rm HWHM} \approx \frac{1.7}{\mu} \tag{15}$$

whereas for a circular lens

$$\left(\frac{\Delta\nu}{\nu_{\rm c}}\right)_{\rm HWHM} \approx \frac{1.5}{\mu^{1/2}} \tag{16}$$

When the focal point is not in the centre of the elliptical region, there will be linear terms proportional to $X\Delta\nu$ and $Y\Delta\nu$ that also contribute to the phase. They make the dependence on $\Delta\nu$ even steeper, which means that equation (14) is an upper bound on the frequency width of magnified images.

It is possible for the frequency and position shifts to cancel each other, so that after moving the source by some distance, it is still strongly magnified at a nearby frequency. That leads to a slope of strongest magnification in time–frequency space. Following a derivation similar to that outlined above for the behaviour with frequency and position separately, we find that the slope is related only to the offset of the source from the focal point. For instance, for the case that the source is travelling along the semi-major axis of the lens, that is, along a trajectory $(x_{\rm s}, y_{\rm s}) = (x_{\rm s}(t), 0)$, we find

$$\frac{{\rm d}\nu}{{\rm d}x_{\rm s}} = \frac{\nu_{\rm c}}{2x_{\rm s}} \tag{17}$$

Finally, it is possible that the pulsar is close to the focus of more than one lens at the same time. As the pulsar moves, these multiple focal points lead to multiple images which change intensity and interfere with each other. This can lead to a wide variety of spectral behaviour, and might well be responsible for the multi-peaked structures seen in some panels of Fig. 2.

A lower bound on the relative velocity. Within the perfect lens approximation, one can derive a lower bound on the relative, transverse velocity between the lens and the pulsar from the durations and frequency widths of the lensing events.

The duration of strongly magnified events is predicted to be

$$\Delta t_{\rm HWHM} \approx \frac{0.7R_{\rm I}^2}{\sqrt{\Delta x_{\rm lens}^2 v_x^2 + \Delta y_{\rm lens}^2 v_y^2}} > \frac{0.7R_{\rm I}^2}{v\Delta x_{\rm lens}} \tag{18}$$

where the inequality corresponds to making the most conservative estimate that the lens is effectively one-dimensional, extended in the x direction, and that the full velocity v is directed along it.

In the frequency space, the same event will have its width bounded from above by equation (14):

$$\left(\frac{\Delta\nu}{\nu}\right)_{\rm HWHM} < \frac{1.8R_{\rm I}^2}{\Delta x_{\rm lens}^2} \tag{19}$$

Combining the two limits to eliminate $\Delta x_{\rm lens}/R_{\rm I}$, we find:

$$v > 0.5\frac{R_{\rm I}}{\Delta t_{\rm HWHM}}\left(\frac{\Delta\nu}{\nu}\right)_{\rm HWHM}^{1/2} \tag{20}$$

Lensing consistency checks. Our analysis in terms of lensing was motivated by the facts that the bright pulses before and after eclipse do not look like giant pulses, that they occur at specific orbital phases where dispersion measure fluctuations are large, and that, in most events, the entire profile increases in flux (although not always by the same factor across the profile), with the enhancements lasting for several pulses.

Consistent with the expectations of plasma lensing, in which lenses focus light but arrive with associated 'shadows' where light is scattered away from our direct line of sight, we find that in regions with many strongly lensed pulses, the average flux received is unchanged (see Extended Data Fig. 3). Also consistent with plasma lensing is that the strongly magnified pulses are highly chromatic, often peaking at high or low frequencies, and sometimes showing slopes in frequency time space or interference patterns characteristic of caustics.

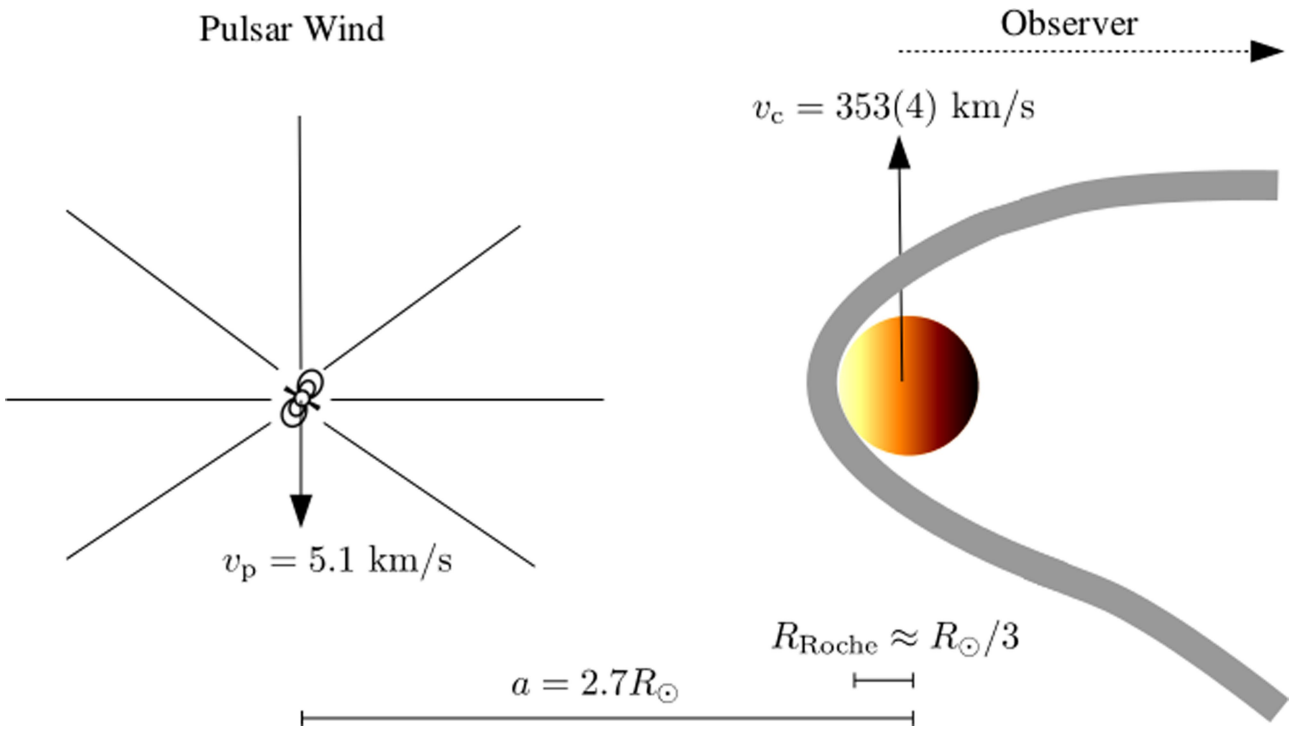
Finally, as a concrete example, the peak magnifications of $\mu \approx 70$ are comparable to the magnifications expected for plasma lensing. The measured dispersive time delay changes provide direct evidence that the geometric and dispersive phases $\phi_{\rm GM}$ and $\phi_{\rm DM}$ can cancel over distances of about 720 km. The cancellation is likely to happen mostly in one spatial direction; setting $\Delta x_{\rm lens} = 720$ km and $\Delta y_{\rm lens} = R_{\rm I}$, our perfect lens model predicts a magnification of the order of 100, comparable to the observed value.

Data availability. The data underlying the figures are available in text files at https://github.com/ramain/B1957LensingData

Code availability. The raw data were read using the baseband package: https://github.com/mhvk/baseband

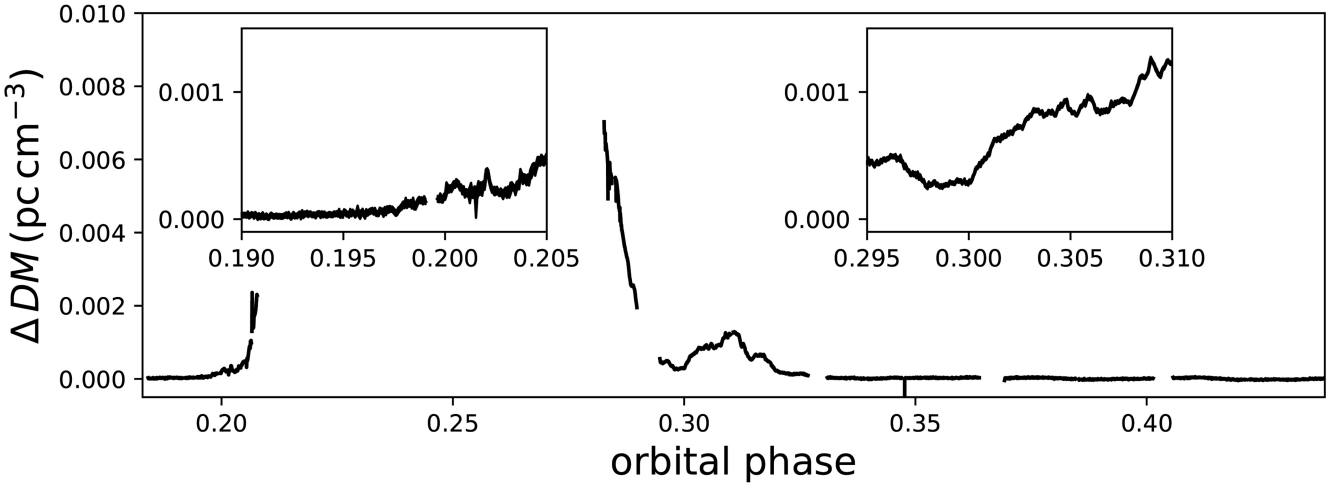
23. Born, M. & Wolf, E. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (Cambridge Univ. Press, Cambridge, 1999).
24. Manchester, R. N. & Taylor, J. H. Parameters of 61 pulsars. *Astrophys. Lett.* **10**, 67–70 (1972).
25. Nye, J. F. *Natural Focusing and Fine Structure of Light: Caustics and Wave Dislocations* (IOP, London, 1999).
26. van Paradijs, J. et al. Optical observations of the eclipsing binary radio pulsar PSR1957 + 20. *Nature* **334**, 684–686 (1988).
27. Reynolds, M. T. et al. The light curve of the companion to PSR B1957+20. *Mon. Not. R. Astron. Soc.* **379**, 1117–1122 (2007).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



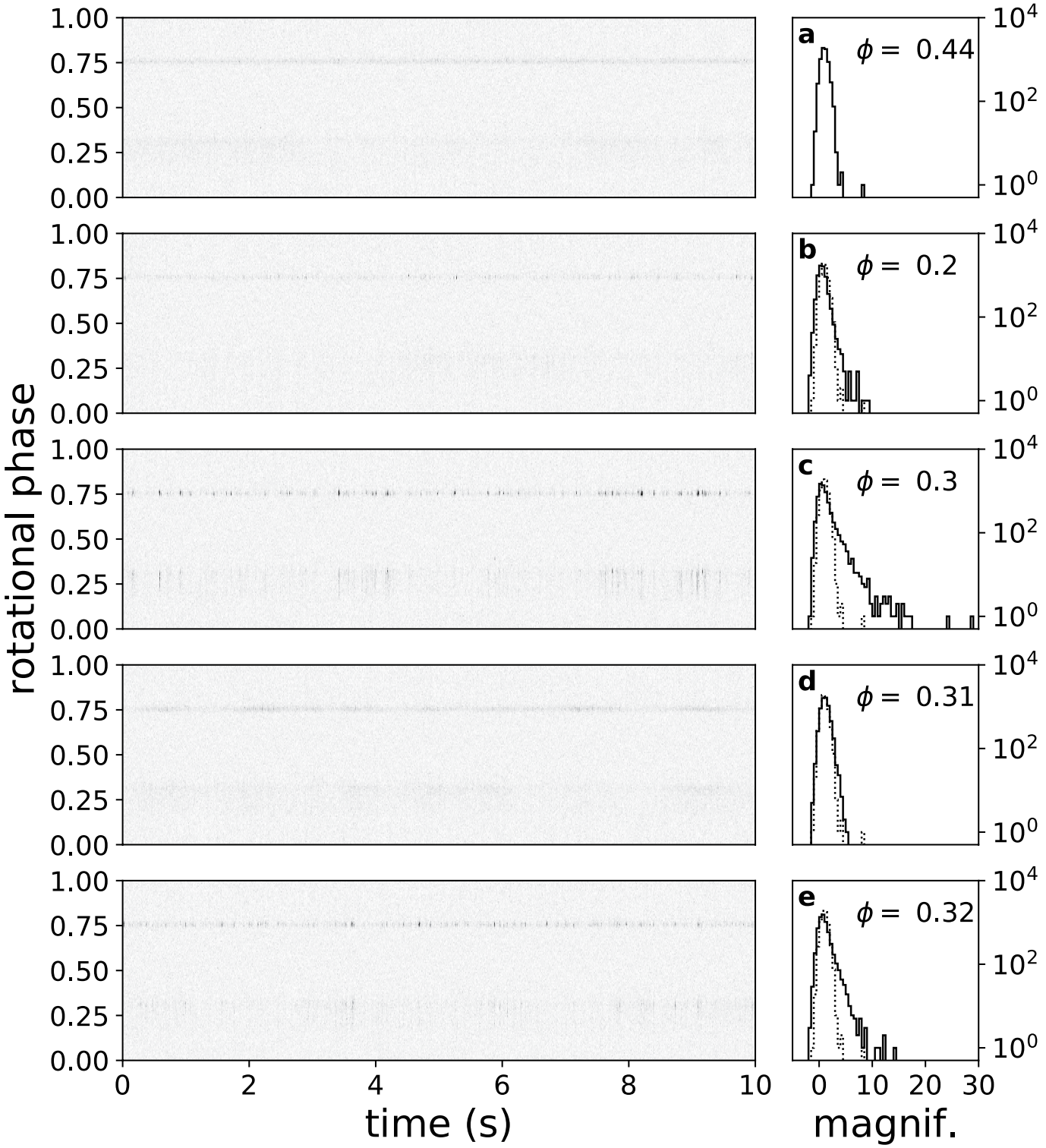
Extended Data Fig. 1 | Shock geometry. The brown dwarf companion is irradiated by the pulsar wind, causing it to be hotter on the side facing the pulsar[26], and inflated, nearly filling its Roche lobe[27]. Outflowing material is shocked by the pulsar wind, leaving a cometary-like tail of material. This tail is asymmetric because of the companion's orbital motion, which leads to eclipse egress lasting substantially longer than ingress. The companion and separations are drawn roughly to scale, whereas the pulsar is not: the light cylinder radius of 76 km would be indistinguishable on this figure. The inclination of the system is conservatively constrained to $50^\circ < i < 85^\circ$ (ref. [13]). v_p and v_c are the velocities of the pulsar and companion, a is the semi-major axis of the orbit, and R_{roche} is the Roche lobe of the companion.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



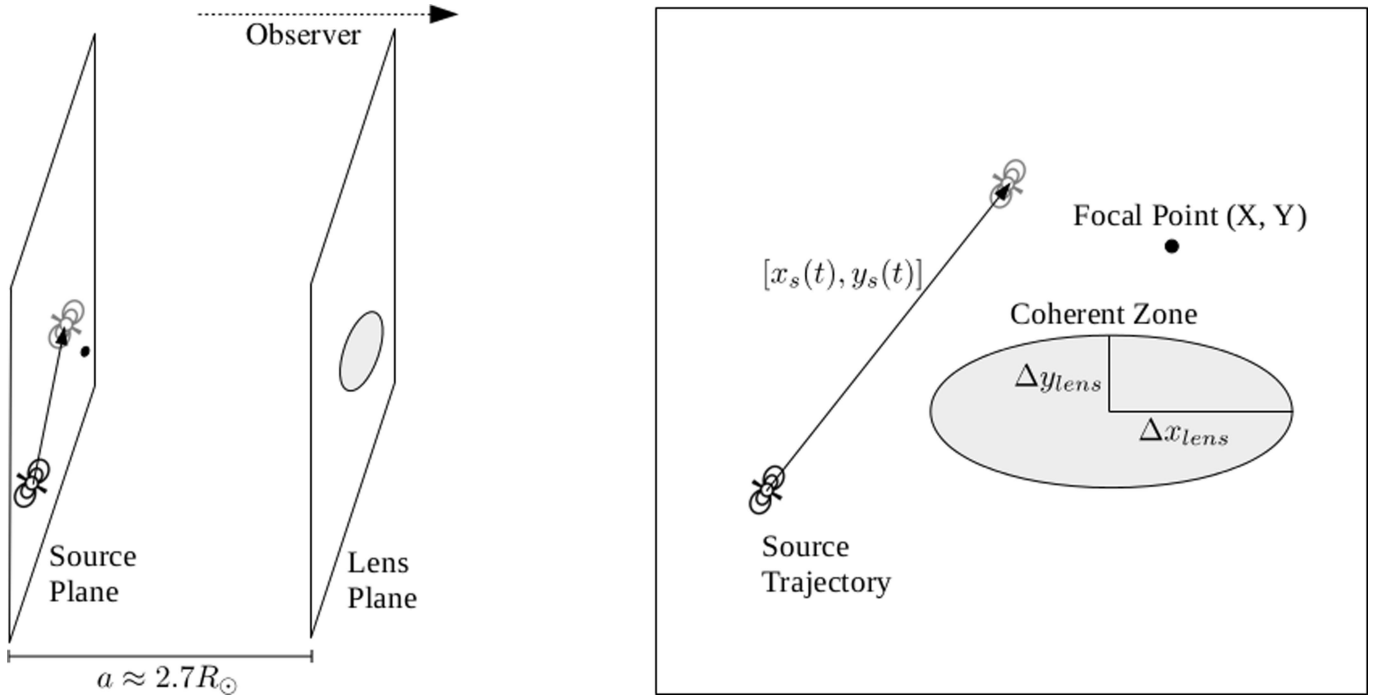
Extended Data Fig. 2 | Dispersion measure near the radio eclipse. Shown is the excess dispersion measure relative to the interstellar dispersion, with insets focusing on regions of strong lensing. The excess dispersion is estimated from delays in pulse arrival times (see Methods); at our observing frequency of 330 MHz, $\Delta\mathrm{DM} = 0.001$ pc cm^{-3} corresponds to a delay of 38 μs. The scatter around the curves is intrinsic. During periods of strong lensing, it is at a level of 2.6×10^{-5} pc cm^{-3}, corresponding to variations in delay time of 1 μs.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 3 | Pulse profiles and magnification distributions. **a**–**e**, Each row shows pulse profiles for 10-s segments (left, in 128 phase bins), and the corresponding magnification distributions. The segments are taken from: **a**, a quiescent region, showing a log-normal magnification distribution (repeated using a dotted line in **b**–**e** for comparison); **b**, eclipse ingress, showing some lensing events, reflected in the tail to high magnification; although hard to see, the average flux is reduced to about 70% by absorption in eclipsing material); **c**, the first post-eclipse lensing period, showing extreme magnifications but little change in average flux; **d**, in between the two post-eclipse strong lensing periods, showing only weak lensing on relatively long, approximately 100-ms timescales; **e**, the second post-eclipse period of strong lensing.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 4 | Geometry of a lensing region. An almost edge-on (left) and a face-on (right) view of the lensing geometry. We assume that the source is at a separation from the pulsar equal to the semi-major axis a of the binary system and moves on a trajectory parallel to the lens plane. A source at the focal point (X, Y) will illuminate the entire elliptical lens coherently, leading to strong magnification. In general, the focal point may not be at the centre of the ellipse (although it is likely to be within it), and the source trajectory $[x_s(t), y_s(t)]$ may not intersect the focal point or the elliptical region.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0101-7

An absolute sodium abundance for a cloud-free 'hot Saturn' exoplanet

N. Nikolov[1]*, D. K. Sing[1,2], J. J. Fortney[3], J. M. Goyal[1], B. Drummond[1], T. M. Evans[1], N. P. Gibson[4], E. J. W. De Mooij[5,6], Z. Rustamkulov[3], H. R. Wakeford[7], B. Smalley[8], A. J. Burgasser[9], C. Hellier[8], Ch. Helling[10,11], N. J. Mayne[1], N. Madhusudhan[12], T. Kataria[13], J. Baines[4], A. L. Carter[1], G. E. Ballester[14], J. K. Barstow[15], J. McCleery[4] & J. J. Spake[1]

Broad absorption signatures from alkali metals, such as the sodium (Na I) and potassium (K I) resonance doublets, have long been predicted in the optical atmospheric spectra of cloud-free irradiated gas giant exoplanets[1–3]. However, observations have revealed only the narrow cores of these features rather than the full pressure-broadened profiles[4–6]. Cloud and haze opacity at the day–night planetary terminator are considered to be responsible for obscuring the absorption-line wings, which hinders constraints on absolute atmospheric abundances[7–9]. Here we report an optical transmission spectrum for the 'hot Saturn' exoplanet WASP-96b obtained with the Very Large Telescope, which exhibits the complete pressure-broadened profile of the sodium absorption feature. The spectrum is in excellent agreement with cloud-free, solar-abundance models assuming chemical equilibrium. We are able to measure a precise, absolute sodium abundance of $\log\varepsilon_{Na} = 6.9^{+0.6}_{-0.4}$, and use it as a proxy for the planet's atmospheric metallicity relative to the solar value ($Z_p/Z_\odot = 2.3^{+8.9}_{-1.7}$). This result is consistent with the mass–metallicity trend observed for Solar System planets and exoplanets[10–12].

We observed two transits of the 'hot Saturn' planet WASP-96b (planetary mass $M_p = (0.48 \pm 0.03)M_J$, where M_J is the mass of Jupiter, planetary radius $R_p = (1.20 \pm 0.06)R_J$, where R_J is the radius of Jupiter, and equilibrium temperature $T_{eq} = 1{,}285 \pm 40$ K)[13] on 2017 July 29 and August 22 UT in photometric conditions, using the 8.2-m Unit Telescope 1 of the Very Large Telescope, with the FORS2 spectrograph. Data were collected in the multi-object-spectroscopy mode using grisms 600B (blue) and 600RI (red) on the first and second nights, respectively, which, when combined, cover the wavelength range 3,600–8,200 Å. We used a mask consisting of two broad slits centred on the target and on a reference star of similar brightness. Broad slits spanning 22″ along the dispersion and 120″ along the spatial (perpendicular) axis were used to minimize slit losses due to seeing variations and guiding imperfections.

For each transit, we produced wavelength-integrated 'white' and spectroscopic light curves for WASP-96 and the reference star by integrating the flux of each spectrum along the dispersion axis. We corrected the light curves for extinction caused by the Earth's atmosphere by dividing the flux of the target by the flux of the reference star. We modelled the transit and systematic effects of the white-light curves by treating the data as a Gaussian process and assuming quadratic limb darkening for the star. The transit parameters—mid-time T_{mid}, orbital inclination i, normalized semi-major axis a/R_*, the planet-to-star radius ratio R_p/R_* and the two limb-darkening coefficients u_1 and u_2—were allowed to vary in the fit to each of the two white-light curves, while the orbital period was held fixed to the previously determined value. The white-light curves and results from the modelling are shown in Extended Data Fig. 1 and Extended Data Table 1.

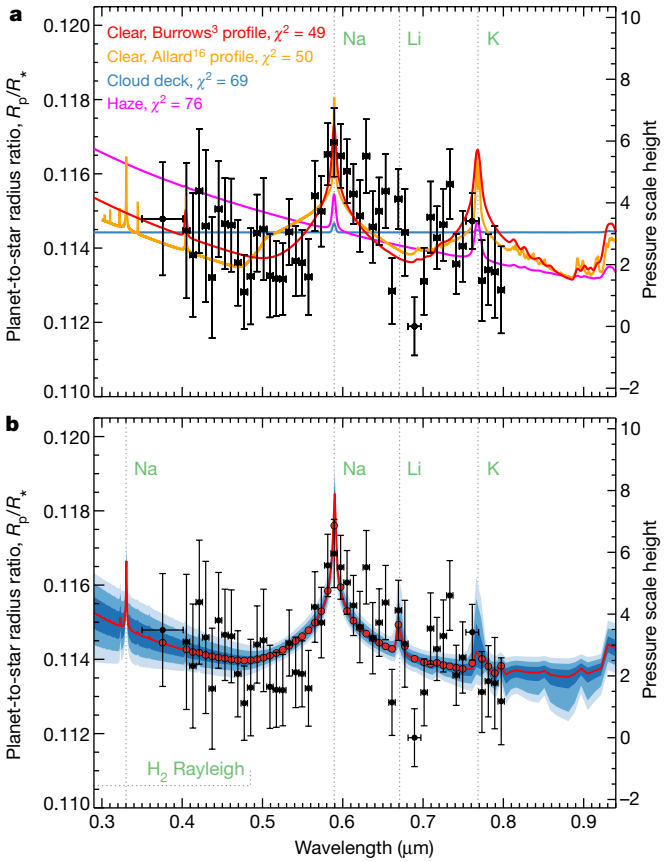


Fig. 1 | Transmission spectrum of WASP-96b compared to models. **a**, Comparison of the FORS2 observations (black dots with 1σ vertical error bars; the horizontal bars indicate spectral bin widths) with clear[3,16], cloudy and hazy one-dimensional forward atmospheric models at solar abundance[14] (continuous lines). The two best-fit models assume a clear atmosphere with different line broadening shapes for Na and K (see text for details). Models with hazes or clouds (magenta and blue) predict much smaller and narrower absorption features. **b**, Similar to **a**, but showing the best-fit model obtained from the retrieval analysis[22] (red line) binned to the data resolution (red dots), with the 1σ, 2σ and 3σ confidence intervals (dark blue to pale blue regions).

[1]Physics and Astronomy, University of Exeter, Exeter, UK. [2]Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Astronomy and Astrophysics, University of California, Santa Cruz, CA, USA. [4]Astrophysics Research Centre, School of Mathematics and Physics, Queens University Belfast, Belfast, UK. [5]School of Physical Sciences, Dublin City University, Glasnevin, Ireland. [6]Centre for Astrophysics & Relativity, Dublin City University, Glasnevin, Ireland. [7]Space Telescope Science Institute, Baltimore, MA, USA. [8]Astrophysics Group, Keele University, Keele, UK. [9]Department of Physics, University of California, San Diego, CA, USA. [10]Centre for Exoplanet Science, SUPA, School of Physics and Astronomy, University of St Andrews, St Andrews, UK. [11]Anton Pannekoek Institute for Astronomy, University of Amsterdam, Amsterdam, The Netherlands. [12]Institute of Astronomy, University of Cambridge, Cambridge, UK. [13]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. [14]Lunar and Planetary Laboratory, University of Arizona, Tucson, AZ, USA. [15]Physics and Astronomy, University College London, London, UK. *e-mail: nikolov.nkn@gmail.com

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

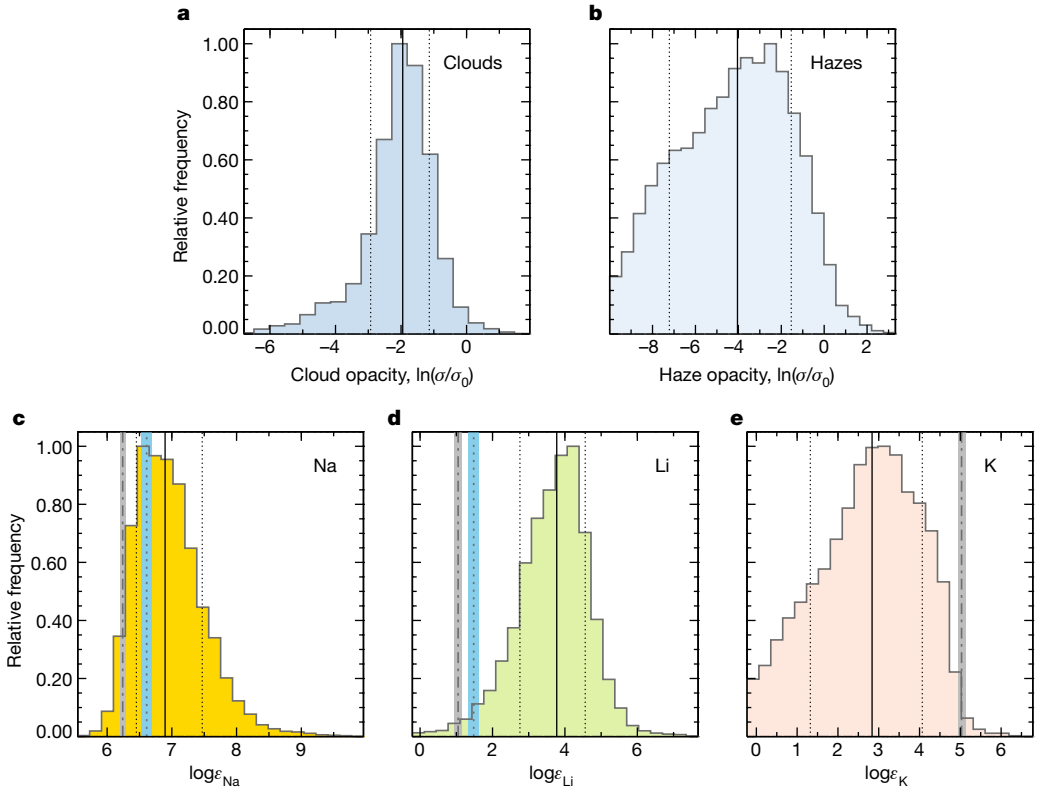


Fig. 2 | Retrieved atmospheric properties for WASP-96b. Histograms of the marginalized posterior distributions from a free retrieval. **a**, **b**, Negligible opacity from clouds and hazes comprises the evidence for a clear atmosphere at the limb of the planet. **c**–**e**, Retrieved elemental abundances in the scale of ref. [23], which ranges from 0 to 12 with the abundance of hydrogen $\log\varepsilon_H = 12$. The abundance of Na, $\log\varepsilon_{Na} = 6.9^{+0.6}_{-0.4}$, is the only constrained quantity (**c**). The vertical continuous and dotted lines indicate the mean abundances and 1σ uncertainties, respectively. Shown are the elemental abundances with the uncertainties of the host star (dotted lines in blue regions) and the Sun (dash-dotted lines in grey regions[25]).

To obtain the transmission spectrum, we produced 28 and 35 spectroscopic light curves, from the blue and red grisms, respectively, with a width of 160 Å. Wavelength-independent systematics were corrected following standard practice, as detailed in the Methods. We allowed only R_p/R_* and u_1 to vary. The rest of the system parameters were fixed to their weighted mean values from the analysis of the two white-light curves, while the quadratic limb-darkening coefficients u_2 were fixed to their theoretical values. To account for systematics, we marginalized over a grid of polynomials, where the latter consisted of terms up to second order in air mass and drift of the spectrum across the detector along both the dispersion and cross-dispersion axes. The resulting time series are shown in Extended Data Figs. 2 and 3.

The measured wavelength-dependent relative planet radii are shown in Fig. 1, which comprises the transmission spectrum of WASP-96b. The spectrum reveals the absorption signature of the pressure-broadened sodium D line with wings covering about 6 atmospheric pressure scale heights (one scale height corresponds to about 610 km, assuming $T_{eq} = 1{,}285$ K), in a wavelength range of around 5,000–7,500 Å, and a slope at near-ultraviolet wavelengths due to Rayleigh scattering by molecular hydrogen. The radius measurements around the potassium feature show no obvious broadened line wing shape or larger absorption at the line cores.

To interpret the measured transmission spectrum, we first compare it with clear, cloudy and hazy atmospheric models with solar abundances from ref. [14]. We find that cloud-free models assuming chemical equilibrium best fitted the 49 data points, giving $\chi^2 = 49$ and $\chi^2 = 50$ for a total of 48 degrees of freedom. Models with clouds and hazes, that is, 100× enhanced-Rayleigh scattering cross-section (haze) and 100× enhanced wavelength-independent (cloud) opacity, give χ^2 values of 69 and 76 respectively, and are disfavoured at about 3σ and about 5σ confidence, respectively (Fig. 1). Further details are provided in Methods.

The wing shape of atomic absorption lines is a result of the combined contribution of the quantum mechanical (natural), thermal (or Doppler) and collisional (or pressure) broadening mechanisms[15]. Measurements of the shape of pressure-broadened line wings can provide important constraints on the interaction potentials used in the theory of stellar and sub-stellar atmospheres[16,17]. Although such constraints have been obtained from Na and K absorption lines in the spectra of brown dwarfs[18,19], the actual shape of the profiles for exoplanets remains unconstrained. To assess the detection of sodium line-broadening we compared the spectrum to models with no broadened lines. Compared to the best-fit clear-atmosphere model, the narrow-line model is found to be rejected at the 5.8σ confidence level. This is in contrast to WASP-39b, WASP-17b and HD209458b, which have previously been classified as having the clearest atmospheres of the known exoplanets. The latter transmission spectra are well explained with narrow alkali features, implying that the broad absorption wings are masked by clouds and hazes[4,5,20,21].

The broad sodium feature measured for WASP-96b therefore provides a unique opportunity to constrain the pressure-broadened line shape for an exoplanet atmosphere. We compared the observed spectrum to two cloud-free models, assuming alkali line-wing shapes from refs [3,16]. We find each of them to be statistically consistent with the data, although the wing profile of ref. [3] is marginally preferred (Fig. 1, red and orange models).

To further interpret the physical properties of WASP-96b's atmosphere, we performed a retrieval analysis of the data using the one-dimensional radiative–convective ATMO model[22]. We assumed an isothermal atmosphere and allowed the temperature, radius, opacity

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

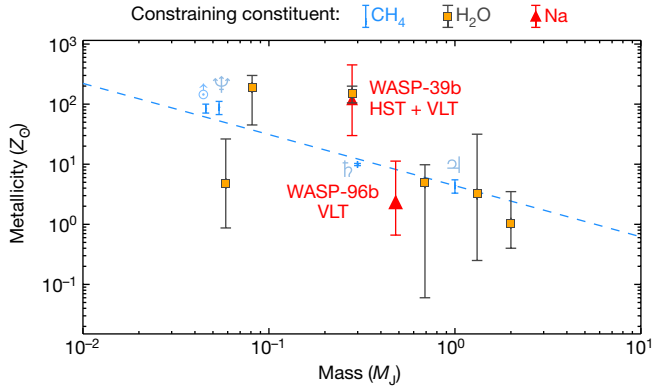


Fig. 3 | Mass–metallicity diagram for Solar System planets and exoplanets. Methane (CH_4) and water (H_2O) are the two absorbing constituents used to constrain the atmospheric metallicity of Solar System planets (blue bars) and hot gas giant exoplanets (orange squares with grey error bars), respectively. Absorption lines from atomic Na (red triangles and error bars) can provide another proxy for exoplanet atmospheric metallicity, by combining three Hubble Space Telescope (HST) and two Very Large Telescope (VLT) transits for WASP-39b. With its detected and resolved pressure-broadened Na line wings, WASP-96b is the first transiting exoplanet for which high-precision atmospheric metallicity has been constrained using data only from the ground. Each error bar corresponds to the 1σ uncertainty. The blue line indicates a fit to the Solar System gas giants (pale blue symbols indicate Solar System planets).

from clouds and hazes and the elemental abundances of Na and K to vary. In addition, Li is expected to add opacity at about 6,650 Å, which is covered by three of our measurements. Throughout this Letter, we adopt the astronomical scale of logarithmic abundances of ref. [23], where hydrogen (H) is defined to be $\log\varepsilon_H = 12$. The abundance of a particular element X is defined as $\log\varepsilon_X = \log(N_X/N_H) + 12$, where N_X and N_H are the number densities of elements X and H. Our retrieval analysis finds negligible contributions from cloud or haze opacity, which indicates that the atmosphere of WASP-96b is free of clouds and hazes at the pressures being probed at the limb. The best-fit transmission spectrum includes opacity from Na, Li, K and Rayleigh scattering (Fig. 1). We obtain a tight constraint of $\log\varepsilon_{Na} = 6.9^{+0.6}_{-0.4}$ on the sodium abundance, which is in agreement with the solar abundance as well as with the measured sodium abundance in the WASP-96 host star (Fig. 2). The best-fit model gives $\chi^2 = 39$ for 42 degrees of freedom.

The current data do not support detections of K or Li, as the minimum χ^2 value when excluding the two species is only slightly higher than when they are included ($\Delta\chi^2 = 2$). However, we include the two species in our retrieval model to marginalize the Na abundance over the possibility of their presence and estimate upper limits on their abundances. The abundances of K and Li are also found to depend on the assumed profile shape of the Na feature. We find an atmospheric temperature of $T = 1{,}710^{+150}_{-200}$ K, which is somewhat higher, compared with the planet's equilibrium temperature of $T_{eq} = 1{,}285 \pm 40$ K under the assumption of zero albedo and uniform day–night heat redistribution[13].

Heavy-element abundance measurements are important to constrain formation mechanisms of gas-giant exoplanets. According to the core-accretion paradigm, as the planet mass decreases, the atmospheric metallicity increases[24,25]. Giant planets accrete H/He-dominated gas as they form, so they also accrete planetesimals[26] that enrich their H/He envelopes in metals. A low-mass H/He envelope has a smaller amount of gas for these metals to be mixed into, leading to a higher metal enrichment compared to the parent star. This is also the scenario for Solar System gas giants, where metallicity has been constrained from methane (CH_4) abundance from in situ or infrared spectroscopy[27–30], showing increasing enrichment of heavy elements with decreasing mass (Fig. 3). Measurements of H_2O abundances have been used to constrain atmospheric metallicities for a small sample of exoplanets[10–12]. The measured molecular abundances are used as proxies to atmospheric metallicities, assuming chemical equilibrium conditions. Using our measurement of the absolute sodium abundance of WASP-96b, we estimate an atmospheric metallicity of $Z_p/Z_\odot = 2.3^{+8.9}_{-1.7}$, that is, $\log(Z_p/Z_\odot) = 0.4^{+0.7}_{-0.5}$. This is consistent with the heavy-element abundance of the host star $Z_*/Z_\odot = 1.4 \pm 0.7$, which we estimate using the relation $Z_*/Z_\odot = 10^{[Fe/H]}$ where $[Fe/H] = 0.14 \pm 0.19$. While our WASP-96b measurement is consistent with the Solar System mass-metallicity trend (see Fig. 3), we note that additional high-precision constraints would be necessary to further support or refute a trend for exoplanets.

WASP-96b is the first exoplanet for which the pressure-broadened wings of an atomic absorption line (Na I) have been observed, probing deeper layers of the atmosphere at the limb. This observation has also enabled a precise atmospheric abundance constraint, using ground-based data alone. Our result demonstrates that combined with near-ultraviolet data, the Na absorption feature at approximately 5,890 Å is a valuable probe of exoplanet metallicities accessible to ground-based telescopes over a wavelength region largely free of contamination by telluric lines. WASP-96b is the first gas giant of approximately 20 exoplanets so far characterized in transmission, to our knowledge, to have a broad atomic absorption feature detected. This demonstrates the important role a future ground-based optical spectrograph, optimized for transmission spectroscopy, could play. With the clearest atmosphere of any exoplanet characterized so far, WASP-96b will be an important target for the upcoming James Webb Space Telescope.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0101-7.

Received: 27 November 2017; Accepted: 30 January 2018;
Published online 7 May 2018.

1. Seager, S. & Sasselov, D. Theoretical transmission spectra during extrasolar giant planet transits. *Astrophys. J.* **537**, 916–921 (2000).
2. Sudarsky, D. et al. Albedo and reflection spectra of extrasolar giant planets. *Astrophys. J.* **538**, 885–903 (2000).
3. Burrows, A. et al. The near-infrared and optical spectra of methane dwarfs and brown dwarfs. *Astrophys. J.* **531**, 438–446 (2000).
4. Charbonneau, D. et al. Detection of an extrasolar planet atmosphere. *Astrophys. J.* **568**, 377–384 (2002).
5. Sing, D. K. et al. A continuum from clear to cloudy hot-Jupiter exoplanets without primordial water depletion. *Nature* **529**, 59–62 (2016).
6. Wyttenbach, A. et al. Hot exoplanet atmospheres resolved with transit spectroscopy (HEARTS). I. Detection of hot neutral sodium at high altitudes on WASP-49b. *Astron. Astrophys.* **602**, A36 (2017).
7. Fortney, J. J. et al. On the indirect detection of sodium in the atmosphere of the planetary companion to HD 209458. *Astrophys. J.* **589**, 615–622 (2003).
8. Line, M. R. & Parmentier, V. The influence of nonuniform cloud cover on transit transmission spectra. *Astrophys. J.* **820**, 78 (2016).
9. Benneke, B. & Seager, S. Atmospheric retrieval for super-Earths: uniquely constraining the atmospheric composition with transmission spectroscopy. *Astrophys. J.* **753**, 100 (2012).
10. Kreidberg, L. et al. A precise water abundance measurement for the hot Jupiter WASP-43b. *Astrophys. J.* **793**, 27 (2014).
11. Line, M. R. et al. No thermal inversion and a solar water abundance for the hot Jupiter HD 209458b from HST/WFC3 spectroscopy. *Astrophys. J.* **152**, 203 (2016).
12. Wakeford, H. et al. HAT-P-26b: a Neptune-mass exoplanet with a well-constrained heavy element abundance. *Science* **356**, 628–631 (2017).
13. Hellier, C. et al. Transiting hot Jupiters from WASP-South, Euler and TRAPPIST: WASP-95b to WASP-101b. *Mon. Not. R. Astron. Soc.* **440**, 1982–1992 (2014).
14. Fortney, J. J., Lodders, K., Marley, M. S. & Freedman, R. S. A unified theory for the atmospheres of the hot and very hot Jupiters: two classes of irradiated atmospheres. *Astrophys. J.* **678**, 1419–1435 (2008).
15. Collins, G. W. *The Fundamentals of Stellar Astrophysics* (W. H. Freeman and Co., New York, 1989).
16. Allard, N. F. et al. A new model for brown dwarf spectra including accurate unified line shape theory for the Na I and K I resonance line profiles. *Astron. Astrophys.* **411**, 473–476 (2003).
17. Burrows, A. & Volobuyev, M. Calculations of the far-wing line profiles of sodium and potassium in the atmospheres of substellar-mass objects. *Astrophys. J.* **583**, 985–995 (2003).
18. Johnas, C. M. S. The effects of new Na I D line profiles in cool atmospheres. *Astron. Astrophys.* **466**, 323–325 (2007).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

19. Burgasser, A. J. The spectra of T dwarfs. II. Red optical data. *Astrophys. J.* **594**, 510–524 (2003).
20. Nikolov, N. et al. VLT FORS2 comparative transmission spectroscopy: detection of Na in the atmosphere of WASP-39b from the ground. *Astrophys. J.* **832**, 191 (2016).
21. Helling, Ch. et al. The mineral clouds on HD 209458b and HD 189733b. *Mon. Not. R. Astron. Soc.* **460**, 855–883 (2016).
22. Tremblin, P. Fingering convection and cloudless models for cool brown dwarf atmospheres. *Astrophys. J.* **804**, 17 (2015).
23. Asplund, M. et al. The chemical composition of the Sun. *Annu. Rev. Astron. Astrophys.* **47**, 481–522 (2009).
24. Fortney, J. J. et al. A framework for characterizing the atmospheres of low-mass low-density transiting planets. *Astrophys. J.* **775**, 80 (2013).
25. Pollack, J. B. et al. Formation of the giant planets by concurrent accretion of solids and gas. *Icarus* **124**, 62–85 (1996).
26. Mordasini, C. et al. Extrasolar planet population synthesis. IV. Correlations with disk metallicity, mass, and lifetime. *Astron. Astrophys.* **541**, A97 (2012).
27. Wong, M. H. et al. Updated Galileo probe mass spectrometer measurements of carbon, oxygen, nitrogen, and sulfur on Jupiter. *Icarus* **171**, 153–170 (2004).
28. Fletcher, L. N. et al. Methane and its isotopologues on Saturn from Cassini/CIRS observations. *Icarus* **199**, 351–367 (2009).
29. Karkoschka, E. & Tomasko, M. G. The haze and methane distributions on Neptune from HST-STIS spectroscopy. *Icarus* **211**, 780–797 (2011).
30. Sromovsky, L. A. et al. Methane on Uranus: the case for a compact CH_4 cloud layer at low latitudes and a severe CH_4 depletion at high-latitudes based on re-analysis of Voyager occultation measurements and STIS spectroscopy. *Icarus* **215**, 292–312 (2011).

Acknowledgements This work is based on observations collected at the European Organization for Astronomical Research in the Southern Hemisphere under European Southern Observatory programme 199.C-0467(H). The research leading to these results received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement number 336792. A.J.B. is a US/UK Fulbright Scholar. J.M.G. and N.J.M. acknowledge support from a Leverhulme Trust Research Project Grant. J.K.B. is a Royal Astronomical Society Research Fellow.

Reviewer information *Nature* thanks I. Snellen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions N.N. led the design of the VLT FORS2 Large Programme and the scientific proposal (with contributions from N.P.G., D.K.S. and T.M.E.). N.N. led the observations, analysis, comparison with forward models and the interpretation of the result. D.K.S. led the atmospheric retrievals (with contributions from J.M.G.). J.J.F. and Z.R. provided forward atmospheric models for comparative analysis. B.S. and C.H. provided elemental abundances of the host star. N.N. wrote the manuscript (with contributions from D.K.S. and T.M.E.). J.B. and J.McC. performed independent tests on various parts of the data reduction and analysis as part of their final-year undergraduate projects under supervision from N.P.G. All authors discussed the results and commented on the draft.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0101-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to N.N.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Observations. We observed two transits of WASP-96b with the FOcal Reducer and Spectrograph (FORS2)[31] attached on the Unit Telescope 1 (Antu) of the VLT at the European Southern Observatory on Cerro Paranal in Chile as part of Large Program 199.C-0467 (Principal Investigator N.N.). We used an observing setup and strategy similar to our VLT FORS2 Comparative Transmission Spectroscopy of WASP-39b and WASP-31b[22,32].

During the two transits, we monitored the flux of WASP-96 and one reference star at photometric conditions. The reference star, known as 2MASS 00041885-4716309, is the only bright source in the FORS2 field of view and is located at an angular separation of 5.3′ away from the target. Fortunately, the reference is of similar colour and brightness, which reduced the effect of differential colour extinction. For example, the magnitude differences (target minus reference) from the PPMXL[33] catalogue are $\Delta B = -0.46$, $\Delta R = -0.49$ and $\Delta I = -0.5$. We observed both transits with the same slit mask and the red detector (Massachusetts Institute of Technology), which is a mosaic of two chips. We positioned the instrument field of view such that each detector imaged one source. The field of view was monitored without guiding interruptions during the full observing campaigns. To improve the duty cycle, we made use of the fastest available read-out mode (200 kHz, about 30 s). On both nights, we ensured that the Longitudinal Atmospheric Dispersion Corrector was in its neutral position, that is, inactive.

During the first night, we used the dispersive element GRIS600B (hereafter blue and 600B), which covers the spectral wavelength range $\lambda = 3{,}600$–$6{,}200$ Å at a resolving power of $R \approx \lambda/\Delta\lambda = 600$. The field of view rose from an air mass of 1.43 to 1.08 and was set at an airmass of 1.16. The seeing oscillated around 0.5″ during the first 3.5 h and gradually increased to 1.2″ at the end of the observation. We collected a total of 89 exposures for about 5 h with integration times adjusted between 120 s and 230 s.

During the second night, we exploited the dispersive element GRIS600RI (hereafter red and 600RI), which covers the range 5,400–8,200 Å, in combination with the GG435 filter, to isolate the first order. The field of view rose from an air mass of 1.23 to 1.08 and was set at an airmass of 1.36. The seeing varied between 0.3″ and 0.5″ as measured from the cross-dispersion profiles of the spectra. We monitored WASP-96 and the reference star for about 5 h 20 min and collected a total of 233 spectra with integration times between 30 s and 80 s.

Calibrations and data reduction. We performed data reduction and analysis using a customized Interactive Data Language (IDL) pipeline[22]. We started by subtracting a bias frame and by applying a flat field correction to the raw images. We computed a master bias and flat field by obtaining the median of 100 individual frames. Cosmic rays were identified and corrected following the routine detailed in ref. [34]. We extracted one-dimensional spectra using the Image Reduction and Analysis Facility (IRAF)'s APALL task. To trace the stars, we used a fit of a Chebyshev polynomial of two parameters. We performed background correction by subtracting the median background from the stellar spectrum for each wavelength, computed from a box located away from the spectral trace. We found that aperture radii of 14 and 12 pixels and sky regions 21 to 72 pixels (where the zero point is the middle of the spectrum special profile) and 23 to 74 pixels minimize the dispersion of the out-of-transit flux of the band-integrated white light curves for the blue and red observations, respectively.

We performed a wavelength calibration of the extracted stellar spectra using spectra of an emission lamp, obtained after each of the two transit observations with a mask identical to the science mask, but with slit widths of 1″. We established a wavelength solution for each of the two stars with a low-order Chebyshev polynomial fit to the centres of a dozen lines, which we identified by performing a Gaussian fit. To account for displacements during the course of each observation and relative to the reference star, we placed the extracted spectra on a common Doppler-corrected rest frame through cross-correlation. All spectra were found to drift in the dispersion direction to no more than 2.5 pixels, with instrument gravity flexure being the most likely reason.

Example spectra of WASP-96 and the reference star are shown in Extended Data Fig. 1. We achieved typical signal-to-noise ratios of 315 and 280 per pixel for the central wavelength of the blue grism and 313 and 257 for the red grism, respectively. We then used the extracted spectra to produce band-integrated white and spectroscopic light curves for each source and transit by summing up the flux along the dispersion axis in each bandpass.

White-light curve analysis. We produced white-light curves from 4,013 Å to 6,173 Å and from 5,293 Å to 8,333 Å for the blue and red observations, respectively. We corrected the raw flux of the target by dividing by the raw flux of the reference star. This correction removes the contribution of Earth's atmospheric transparency variations, as demonstrated in Extended Data Fig. 1. We modelled the white light transits and instrumental systematics simultaneously by treating the data as a Gaussian process[35–37]. We performed the Gaussian process analysis using the Python Gaussian process library *George*[38–41]. Under the Gaussian process assumption, the data likelihood is a multivariate normal distribution with a mean function μ describing the deterministic transit signal and a covariance matrix K that accounts for stochastic correlations (that is, poorly constrained systematics) in the data:

$$p(\boldsymbol{f} \mid \boldsymbol{\theta}, \gamma) = \mathcal{N}(\mu, K)$$

where p is the probability density function, $\boldsymbol{f}$ is a vector containing the flux measurements, $\boldsymbol{\theta}$ is a vector containing the mean function parameters, γ is a function containing the covariance parameters and $\mathcal{N}$ is a multivariate normal distribution. We defined the mean function μ as follows:

$$\mu(\boldsymbol{t}, \hat{\boldsymbol{t}}; c_0, c_1, \boldsymbol{\theta}) = [c_0 + c_1\hat{\boldsymbol{t}}]\,T(\boldsymbol{t}; \boldsymbol{\theta})$$

where $\boldsymbol{t}$ is a vector of all central exposure time stamps in Julian Date, $\hat{\boldsymbol{t}}$ is a vector containing all standardized times, that is, with subtracted mean exposure time and divided by the standard deviation, c_0 and c_1 describe a linear baseline trend, $T(\theta)$ is an analytical expression describing the transit and $\boldsymbol{\theta} = (i, a/R_*, T_{\rm mid}, R_{\rm p}/R_*, u_1, u_2)$, where i is the orbital inclination, a/R_* is the normalized semi-major axis, $T_{\rm mid}$ is the central transit time, $R_{\rm p}/R_*$ is the planet-to-star radius ratio, and u_1 and u_2 are the linear and quadratic limb darkening coefficients. To obtain an analytical transit model $T(\boldsymbol{\theta})$, we used the formulae found in ref. [42]. We fixed the orbital period to its value from ref. [13] and fitted for the remaining system parameters.

We accounted for the stellar limb-darkening by adopting the two-parameter (u_1, u_2) quadratic law and computed the values of the coefficients using a three-dimensional stellar atmosphere model grid[43]. In these calculations, we adopted the closest match to the effective temperature, surface gravity and metallicity of the exoplanet host star found in ref. [13]. The choice of a quadratic versus a more complex law (such as a four-parameter nonlinear law[44]) was motivated by the study of refs [45,46], in which the two-parameter law has been shown to introduce negligible bias on the measured properties of transiting systems similar to WASP-96. In addition, the quadratic law requires a much shorter computational time to determine the relevant transit light curve. We computed the theoretical limb-darkening by fitting the limb-darkened intensities of the three-dimensional stellar atmosphere models, factored by the throughputs of the blue and red grisms.

The covariance matrix is defined as $K = \sigma_i^2\delta_{ij} + k_{ij}$, where σ_i are the photon noise uncertainties, δ_{ij} is the Kronecker delta function and k_{ij} is a covariance function. We assumed the white noise term $\sigma_{\rm w}$ was the same for all data points and allowed it to vary as a free parameter. For the covariance function, we chose to use the Matérn $\nu = 3/2$ kernel with the spectral dispersion and cross-dispersion drifts $\boldsymbol{x}$ and $\boldsymbol{y}$ as input variables, and the full-width at half-maximum (FWHM) measured from the cross-dispersion profiles of the two-dimensional spectra and the speed of the rotation angle $\boldsymbol{z}$ (see Extended Data Fig. 4). As with the linear time term, we also standardized the input parameters before the light curve fitting. We chose to use the dispersion and cross-dispersion drifts for both observations, and combined them with the FWHM for the blue data and the speed of the rotation angle for the red data, respectively. Our choice was justified based on the fact that those combinations of input parameters gave well behaved residuals. The covariance function was then defined as:

$$k_{ij} = A^2(1 + \sqrt{3}\,D_{ij})\exp(-\sqrt{3}\,D_{ij})$$

where A is the characteristic correlation amplitude and

$$D_{ij} = \sqrt{\frac{(\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_j)^2}{\tau_x^2} + \frac{(\hat{\boldsymbol{y}}_i - \hat{\boldsymbol{y}}_j)^2}{\tau_y^2} + \frac{(\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{z}}_j)^2}{\tau_z^2}}$$

where τ_x, τ_y and τ_z are the correlation length scales and the hatted variables are standardized. We allowed parameters $\boldsymbol{X} = (c_0, c_1, T_{\rm mid}, i, a/R_*, R_{\rm p}/R_*, u_1, u_2)$ and $\boldsymbol{Y} = (A, \tau_x, \tau_y, \tau_z)$ to vary and fixed the orbital period P to its literature value[13]. We adopted uniform priors for $\boldsymbol{X}$ and log-uniform priors for $\boldsymbol{Y}$.

To marginalize the posterior distribution $p(\theta, \gamma \mid f) \propto p(f \mid \theta, \gamma)p(\theta, \gamma)$ we made use of the Markov chain Monte Carlo software package *emcee*[40]. We identified the maximum likelihood solution using the Levenberg–Marquardt least-squares algorithm[47] and initialized three groups of 150 walkers close to that maximum. We run groups one and two for 350 samples and the third group had 4,500 samples. Before running for the second group we re-sampled the positions of the walkers in a narrow space around the position of the best walker from the first run. This extra re-sampling step was useful because otherwise some of the walkers can start in a low-likelihood area of parameter space and require more computational time to converge. Transit models for each of the two observations computed using the marginalized posterior distributions are shown in Extended Data Fig. 1 and the relevant parameter values are reported in Extended Data Table 1. We find residual dispersion of 79 and 203 parts per million for the blue and red light curves, respectively. Both values are found to be within 80% of the theoretical photon noise limit.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

We computed the weighted mean values of the orbital inclination and semi-major axis and repeated the fits. In the second fit we allowed only the planet-to-star radius ratio R_p/R_* and the two limb darkening coefficients u_1 and u_2 to vary, while the orbital inclination and semi-major axis were held fixed to the weighted mean values and the central times were fixed to the values determined from the first fit.

Spectroscopic light curve analysis. We produced spectroscopic light curves by summing the flux of the target and reference star in bands with a width of 160 Å. The sodium D lines at 5,890 Å and 5,896 Å fall inside the spectral range of each grism, within their overlapping region of 5,300–6,200 Å. We centred the set of bins for each night on the sodium line, which gave the advantage of obtaining two radius measurements identical in wavelength coverage within that overlapping region. We merged two pairs of bins, covering the O_2 A and B bands from 7,594 Å to 7,621 Å and from 6,867 Å to 6,884 Å, respectively, to increase the signal-to-noise ratio of the corresponding light curves. The very first band in the blue grism was also enlarged with the same motivation. With these customizations we produced a total of 63 light curves.

Common mode factors. The FORS2 spectroscopic light curves are known to exhibit wavelength-independent (common mode) systematic effects, as demonstrated from our Comparative Transmission Spectroscopy studies[22,32] and other FORS2 results[48–50]. This makes the instrument outstanding for transmission spectroscopy, with enormous potential to explore the diversity of exoplanet atmospheres. We established the wavelength-independent systematic effects using the band-integrated white-light curves for each of the two nights. We simply divided the white-light transit light curves by a transit model. We computed the transit model using the weighted mean values of the orbital inclination and normalized semi-major axis from both nights and assumed the central times found from the white-light analysis. The values for the relative radius and the limb-darkening coefficients were identified by repeating the Gaussian process fit where the transit central time, orbital inclination and semi-major axis were held fixed to the weighted-mean values. The fitted relative radii and limb-darkening coefficients are reported at the end of Extended Data Table 1. The common mode factors for each night are shown in Extended Data Fig. 1 along with a schematic explanation of the full white-light curve analysis.

Spectroscopic light curve fits. We modelled the spectroscopic light curves using a two-component function that takes into account the systematics and transit simultaneously. The transit model was computed using the analytical formulae of ref. [42], as for the white-light curve, but in the fits we allowed only the relative planet radius R_p/R_* and the linear limb-darkening coefficient u_1 to vary. Fitting only for the linear limb-darkening coefficient is standard practice for ground-based observations and has proved generally to perform well[51–56]. Similar to our WASP-39b study with FORS2[22], we also fitted for both limb-darkening coefficients and found that the uncertainty of u_2 is large and consistent with the theoretical prediction. We interpret this as an indication for insufficient constraining power of the data for the nonlinear coefficient. However, as the transmission spectra did not substantially change we chose to fix u_2 and to fit only for u_1. Prior to fitting the spectroscopic light curves, we removed the common mode factors from each night by dividing each of the spectroscopic light curves by the corresponding common-mode light curve of the same night. We accounted for the systematics using a low-order polynomial (up to second degree with no cross terms) of dispersion and cross-dispersion drift, air mass, FWHM variations and the rate of change of the rotator angle (Extended Data Fig. 4). We produced all possible combinations of detrending variables and performed separate fits with each combination with the systematics function included in the two-component model. For each attempted function, we computed the Akaike Information Criterion[57] to estimate the statistical weight of the model depending on the number of degrees of freedom. We marginalized the resulting relative radii and linear limb-darkening coefficient following ref. [58]. We chose to rely on the Akaike Information Criterion instead of other information criteria, for example, the Bayesian Information Criterion[59], because the Akaike Information Criterion selects more complex models, resulting in more conservative error estimates. We note that a marginalization over multiple systematics functions relies on the assumption of equal prior weights for each tested model. This assumption is valid for simple polynomial expansions of basis parameters, like the ones in our study. We found systematics models parameterized with linear air mass, dispersion drift and FWHM terms to result in the highest statistical weight.

Prior to each fit, we set the uncertainties of each spectrophotometric channel to values that are based on the expected photon noise with additional component from readout noise. We determined best-fit models using a Levenberg–Marquardt least-squares algorithm and rescaled the uncertainties of the fitted parameters with the dispersion of the residuals. All residual outliers larger than 3σ were excluded from the analysis. We found that for each spectroscopic light curve fewer than three data points were removed. We also assessed the levels of correlated residual red noise, following the methodology of ref. [60] by modelling the binned variance with the $\sigma^2 = \sigma_w^2/N + \sigma_r^2$ relation, where σ_w is the uncorrelated white noise component, N is the number of measurements in the bin and σ_r is the red noise component. We find white and red noise dispersion in the range 400–1,000 and 20–80 parts per million, respectively.

The measured wavelength-dependent relative radii and corresponding light curves are given in Fig. 1, Extended Data Figs. 2–5, and Extended Data Table 2 and comprise the transmission spectrum of WASP-96b. We used the overlapping wavelength region to combine the two observations and computed the weighted mean of both datasets. We detect a marginally significant 1.4σ difference in the transit depths of the light curves from the two observations of $(7.2 \pm 5.2) \times 10^{-4}$. This level of variation is consistent with the photometric variability, associated with the active regions on the star surface with variability of $\sigma < 9.2 \times 10^{-4}$ (ref. [13]).

Modelling of the atmosphere of WASP-96b. First we compare the observed transmission spectrum to the models based on ref. [16], which included a self-consistent treatment of radiative transfer and chemical equilibrium of neutral and ionic species. Chemical mixing ratios and opacities were computed assuming solar metallicity and local chemical equilibrium, accounting for condensation and thermal ionization but not photoionization[61–63]. Transmission spectra were also calculated using one-dimensional T-P profiles for the dayside, as well as an overall cooler planetary-averaged profile.

In addition to the atmospheric models of a clear atmosphere, a simplified treatment of the scattering and absorption have been incorporated in those models to simulate the effect of small particle haze aerosols and large particle cloud condensates at optical and near-infrared wavelengths. In the case of haze, Rayleigh scattering opacity ($\sigma = \sigma_0(\lambda/\lambda_0)^{-4}$) has been assumed with a cross-section which was 1,000 × the cross-section of molecular hydrogen gas ($\sigma_0 = 5.31 \times 10^{-27}$ cm^2 at $\lambda_0 = 3{,}500$ Å; ref. [61]). To include the effects of a flat cloud deck we included a wavelength-independent cross-section, which was a factor of 100× the cross-section of molecular hydrogen gas at $\lambda_0 = 3{,}500$ Å (see Fig. 1).

As in our previous studies, we obtained the average values of the models within the wavelength bins and fitted these theoretical values to the data with a single parameter responsible for their vertical offset[5,22,34]. The χ^2 and Bayesian Information Criterion statistic quantities were computed for each model with the number of degrees of freedom for each model determined by $\nu = N - m$, where N is the number of measurements and m is the number of free parameters in the fit.

We also compared the observed transmission spectrum around the Na D line from 4,500 Å to 6,800 Å with cloud-free atmosphere models, assuming two individual shapes of the pressure broadened profile. The first profile had the shape detailed in ref. [3] and the second line shape was taken from the formalism of refs [18,26]. We find no statistically significant difference between the two with $\chi^2 = 22$ and $\chi^2 = 30$ for 35 degrees of freedom. Further observations at higher signal-to-noise ratios and resolution will be necessary to distinguish between the two line-wing shapes.

Retrieval analysis. We performed a retrieval[8,9,64] analysis using the one-dimensional radiative–convective equilibrium ATMO model[24,65–68]. The code includes isotropic multi-gas Rayleigh scattering, H_2–H_2 and H_2–He collision-induced absorption, as well as opacities for all major chemical species taken from the most up-to-date high-temperature sources, including H_2, He, H_2O, CO_2, CO, CH_4, Na, K, Li, Rb and Cs[66,69]. Hazes and clouds are respectively treated as parameterized enhanced Rayleigh scattering and grey opacity[70], which is similar to the approach of ref. [16] and is detailed in ref. [69]. ATMO uses the correlated-k approximation with the random overlap method to compute the total gaseous mixture opacity, which has been shown to agree well with a full line-by-line treatment[66].

We first performed a free retrieval analysis, allowing a fit to the abundances of Na, K and Li, cloud and haze opacity, the planet's radius and atmospheric temperature. Na, K and Li were included because these elements are expected to add substantial opacity in the wavelength region of our observations. We assumed that these gases are well mixed vertically in the atmosphere.

For all free parameters in our model we adopted uniform priors with the following ranges: 250–3,000 K for the temperature, $1R_J$–$1.5R_J$ for the radius, 10^{-10} to 10^3 for the cloud and 10^{-10} to 10^3 for the haze opacity. For the mixing ratios of chemical species other than H and He we adopted uniform priors between −2 and 12. Our retrieval analysis proceeded by first identifying the minimum χ^2 solution using nonlinear least-squares optimization and then marginalizing over the posterior distribution using differential-evolution Markov chain Monte Carlo[71]. A total of 12 chains for 30,000 steps each were run until the Gelman–Rubin statistic for each free parameter was within 1% of unity, showing that the chains were well mixed and had reached a steady state. We discarded a burn-in phase from all chains corresponding to the step at which all chains had found a χ^2 below the median χ^2 value of the chain. Finally, we combined the remaining samples into a single chain, forming our posterior distributions. We summarize the results from the free retrieval in Fig. 2.

Elemental abundances of Na and Li were obtained for the host star from high-resolution optical spectroscopy with a typical signal-to-noise ratio of about 100:1, as detailed in ref. [13]. Because the K resonance doublet is located outside the red limit of the high-resolution spectra, it was not possible to measure its abundance.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

The best-fitting retrieved spectrum suggests a cloud-free atmosphere at the limb with precise sodium abundance consistent with the solar value at about 1σ. The spectrum shows no definitive evidence of the potassium feature. A measurement at around 0.73 μm shows a slightly larger radius, though it is only about 1.8σ above our best-fitting model. Our retrieval analysis shows a marginalized posterior distribution with only an upper bound with median value consistent with the solar value at approximately 2σ confidence (Figs. 1 and 2). This is most probably a consequence of the limited wavelength range, covering only about two-thirds of the potassium feature, as well as the diluting effect of the O_2-A band, which partially covers this wavelength regime.

The limited constraining power of the K feature allows for a three-case scenario for the abundance of K in the atmosphere of WASP-96b. In all three cases the atmosphere of the planet is clear at the limb with sodium at solar abundance. In the first scenario, the abundance of potassium is sub-solar, leading to a missing potassium feature in the transmission spectrum. In chemical equilibrium, Na and K can form condensates (such as Na_2S and KCl) at temperatures lower than about 1,300 K and 1,000 K, respectively. However, given that the retrieved temperature in the region probed by our observation lies between about 1,300 K and 1,700 K, this is unlikely. In the second scenario, the potassium can be ionized by the ultraviolet radiation of the host star, which would also lead to missing potassium line cores. The pressure-broadened line wings would still be present, because they would originate from deeper layers, where the ultraviolet radiation would be able to penetrate. We estimate the relative sodium-to-potassium ratio from our free retrieval analysis, finding a highly uncertain value, which prevents any definitive conclusion (Extended Data Fig. 6). Future observations of multiple transits could help resolve the two-case scenario. A third scenario would require a stratified atmosphere consisting of a low-altitude potassium layer with cloud cover on top, above which is located a clear atmosphere containing sodium at solar abundance. The sodium layer would need to be deep enough to allow for pressure-broadened line wings in the transmission spectrum.

We note a large degeneracy between the aerosol clouds/hazes and the Na abundance, as higher cloud opacity levels can be fitted with increased Na abundances. This degeneracy can be seen in the posterior distribution of the retrieval (see Extended Data Fig. 8), and the marginalized posterior distributions are shown in Fig. 2 (that is, the distribution shown in Extended Data Fig. 8 integrated along each axis). The degeneracy is accounted for in our retrieval modelling by marginalizing the Na abundance over the other model fit parameters, which includes the effects of clouds and hazes. An aerosol-free model where the near-ultraviolet transmission spectra is dominated by H_2 Rayleigh scattering helps us to determine the lower limit to the Na abundance. The upper limit for the Na abundance is sensitive to the line profile shape, as very high aerosol opacity levels probe much lower pressure levels, affecting the line profile. For example, for a clear atmosphere the transmission spectra at 4,000 Å probes pressures of 20 mbar and is dominated by molecular hydrogen at that wavelength. In a model where the haze opacity is 1,000× stronger than molecular hydrogen (the pink haze model shown in Fig. 1), the pressure probed is about 1,000× lower or about 0.02 mbar. These lower pressures affect the calculation of the pressure-broadened sodium line profile. As described in ref. [3], using semi-classical impact theory the half-width of the Lorentzian sodium line core is determined by the effective collision frequency, which itself is a product of the H_2-perturber density (among other factors). We find that in the described scenario with very high sodium abundances and high cloud/haze opacities, very low pressures (such as 0.02 mbar) are probed and the sodium line profile becomes too narrow to be able to fit the wide profile seen in the data, even at arbitrarily high abundances.

We acknowledge that the one-dimensional model does not account for horizontal advection, which may affect the composition. This can break the chemical equilibrium condition, leading to horizontally constant non-equilibrium abundances of various species for example, CH_4 (ref. [72]), which could potentially affect the observed transmission spectrum. However, an analysis of such effects is beyond the scope of the present Letter.

The presence of sodium and potassium in the atmospheric spectra of irradiated gas giant exoplanets has proved puzzling, with some of the spectra showing both or either of the two features without a clear trend associated with the planetary properties[5]. Primordial abundance variation along with atmospheric processes such as condensation, photochemistry and photoionization have been hypothesized to be responsible for the observed alkali variation. Elemental abundance constraints with ground-based instruments such as FORS2 could help identify the role of some of those processes from statistically large samples of exoplanet spectra.

Exoplanet atmospheric metallicities have been estimated using the scaling relation of atmospheric metallicity with the abundance of detected absorption features. At present, such estimations have been obtained using the 1.4-μm water absorption feature. Recently, ref. [73] obtained such measurement from a combined HST + VLT spectroscopy. We performed a retrieval analysis assuming chemical equilibrium in the atmosphere and determine the metallicity of WASP-96b's atmosphere, finding consistency with the metallicity of the host star and the mass–metallicity trend established for Solar System giant planets (Fig. 3). We acknowledge that such estimations can largely be inaccurate owing to the metallicity scaling assumption with elemental abundance and round off the precision of the WASP-96b's metallicity to one significant figure. The current census of exoplanet metallicity measurements exhibits a large scatter across the full mass–metallicity diagram, which hampers definitive conclusion regarding a trend. As in ref. [74] we plot the relative planet-to-star heavy-element enrichment as a function of the planet mass instead of the planet metallicity alone (Extended Data Fig. 7). We convert $Z_p/Z_\odot$ to Z_p/Z_*, assuming a scaling approximation with the parent star's iron abundance of the form $Z_p/Z_* = 10^{[Fe/H]}$, where $Z_\odot = 0.014$ and propagate the uncertainty of [Fe/H]. It remains for future work to improve the statistics on the mass–metallicity diagram with additional precise measurements. We therefore consider it premature to propose any trends or emerging patterns or interpretation.

Code availability. Publicly available custom codes were used for the Gaussian process modelling: *George* (http://dfm.io/george/current/user/gp/) and *emcee* (http://github.com/dfm/emcee). The Markov chain Monte Carlo retrieval analyses were performed using the publicly available package *exofast* (http://astroutils.astronomy.ohio-state.edu/exofast). The ATMO code used to compute the atmosphere models is currently proprietary. We have opted not to make the customized IDL codes used to produce the spectra publicly available owing to their undocumented intricacies.

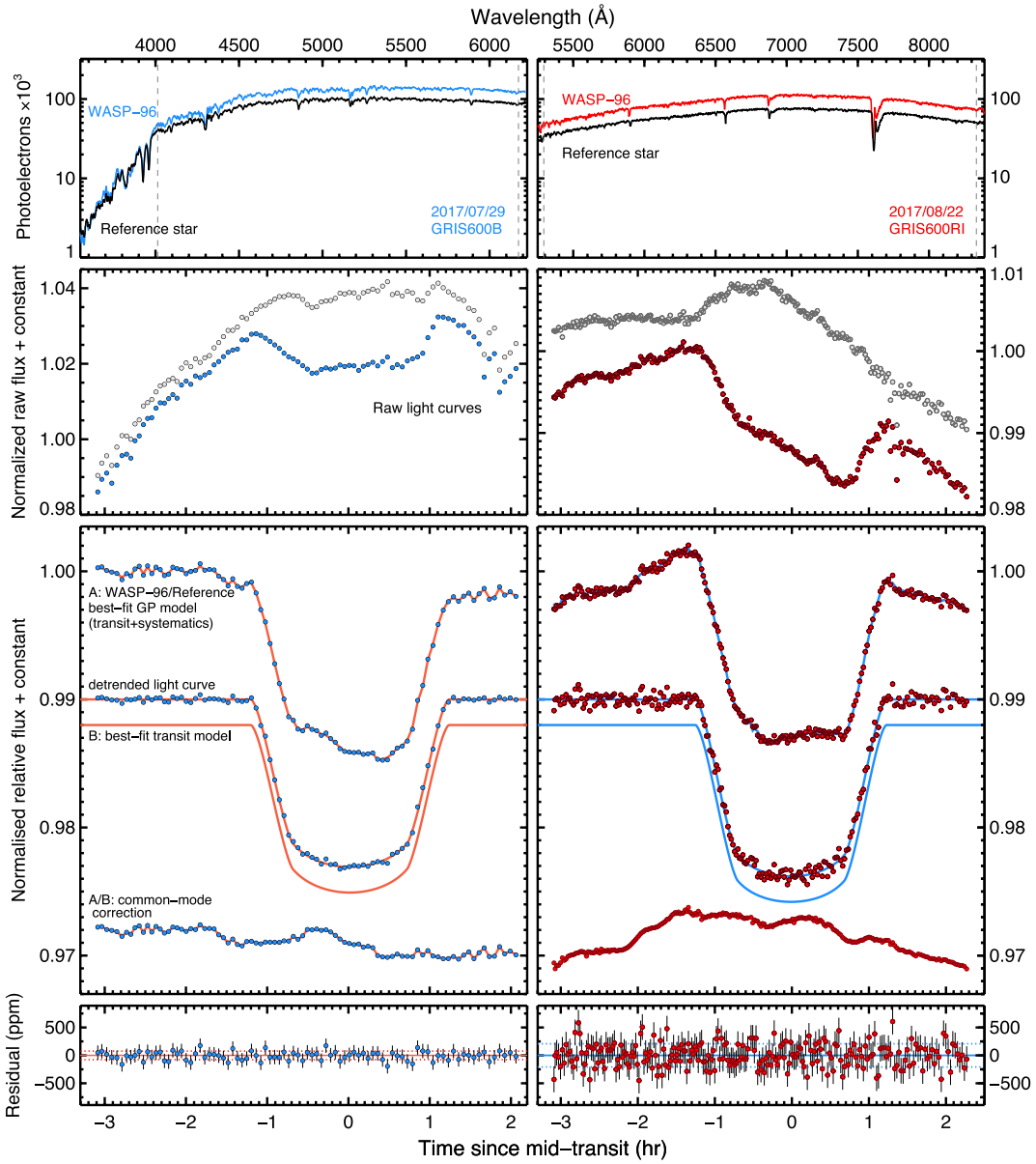
Data availability. The data are within the standard proprietary period of one year and will become publicly available on the European Southern Observatory archive (July and August) in 2018.

31. Appenzeller, I. et al. Successful commissioning of FORS1—the first optical instrument on the VLT. *The Messenger* **94**, 1–6 (1998).
32. Gibson, N. P. et al. VLT/FORS2 comparative transmission spectroscopy II: confirmation of a cloud deck and Rayleigh scattering in WASP-31b, but no potassium? *Mon. Not. R. Astron. Soc.* **467**, 4591–4605 (2017).
33. Roeser, S. et al. The PPMXL catalogue of positions and proper motions on the ICRS. Combining USNO-B1.0 and the two micron all sky survey (2MASS). *Astron. J.* **139**, 2440–2447 (2010).
34. Nikolov, N. et al. Hubble Space Telescope hot Jupiter transmission spectral survey: a detection of Na and strong optical absorption in HAT-P-1b. *Mon. Not. R. Astron. Soc.* **437**, 46–66 (2014).
35. Gibson, N. P. et al. A Gaussian process framework for modelling instrumental systematics: application to transmission spectroscopy. *Mon. Not. R. Astron. Soc.* **419**, 2683–2694 (2012).
36. Evans, T. M. et al. An ultrahot gas-giant exoplanet with a stratosphere. *Nature* **548**, 58–61 (2017).
37. Nikolov, N. et al. Hubble PanCET: an isothermal day-side atmosphere for the bloated gas-giant HAT-P-32Ab. *Mon. Not. R. Astron. Soc.* **474**, 1705–1717 (2018).
38. Ambikasaran, S. et al. Fast direct methods for Gaussian processes. *IT Process Automation Manager (ITPAM)* **38**, 252–265 (2015).
39. Foreman-Mackey, D. George: Gaussian process regression. *Astrophys. Source Code Library* ascl.soft11015 (2015).
40. Foreman-Mackey, D. et al. emcee: the MCMC hammer. *Proc. Astron. Soc. Pacif.* **125**, 306–312 (2013).
41. Foreman-Mackey, D. corner.py: scatterplot matrices in Python. *J. Open Source Softw.* **1**, 24 (2016).
42. Mandel, K. & Agol, E. Analytic lightcurves for planetary transit searches. *Astrophys. J.* **580**, L171–L175 (2002).
43. Magic, Z. et al. The Stagger-grid: a grid of 3D stellar atmosphere models. IV. Limb darkening coefficients. *Astron. Astrophys.* **573**, A90 (2015).
44. Claret, A. A new non-linear limb-darkening law for LTE stellar atmosphere models II. Geneva and Walraven systems: calculations for $-5.0 \leq \log[M/H] \leq +1$, 2000 K $\leq T_{eff} \leq$ 50000 K at several surface gravities. *Astron. Astrophys.* **401**, 657–660 (2003).
45. Espinoza, N. & Jordan, A. Limb darkening and exoplanets: testing stellar model atmospheres and identifying biases in transit parameters. *Mon. Not. R. Astron. Soc.* **450**, 1879–1899 (2015).
46. Espinoza, N. & Jordan, A. Limb darkening and exoplanets—II. Choosing the best law for optimal retrieval of transit parameters. *Mon. Not. R. Astron. Soc.* **457**, 3573–3581 (2016).
47. Markwardt, C. B. Non-linear least-squares fitting in IDL with MPFIT. *Astron. Soc. Pacif. Conf. Ser.* **411**, 251–254 (2009).
48. Sedaghati, E. et al. Detection of titanium oxide in the atmosphere of a hot Jupiter. *Nature* **549**, 238–241 (2017).
49. Sedaghati, E. et al. Potassium detection in the clear atmosphere of a hot-Jupiter. FORS2 transmission spectroscopy of WASP-17b. *Astron. Astrophys.* **596**, A47 (2016).
50. Lendl, M. FORS2 observes a multi-epoch transmission spectrum of the hot Saturn-mass exoplanet WASP-49b. *Astron. Astrophys.* **587**, A67 (2016).
51. Southworth, J. Homogeneous studies of transiting extrasolar planets—I. Light-curve analyses. *Mon. Not. R. Astron. Soc.* **386**, 1644–1666 (2008).
52. Mallonn, M. et al. Transmission spectroscopy of the inflated exo-Saturn HAT-P-19b. *Astron. Astrophys.* **580**, A60 (2015).
53. Stevenson, K. B. A search for water in the atmosphere of HAT-P-26b using LDSS-3C. *Astrophys. J.* **817**, 141 (2016).
54. Nikolov, N. et al. HST hot-Jupiter transmission spectral survey: haze in the atmosphere of WASP-6b. *Mon. Not. R. Astron. Soc.* **447**, 463–478 (2015).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

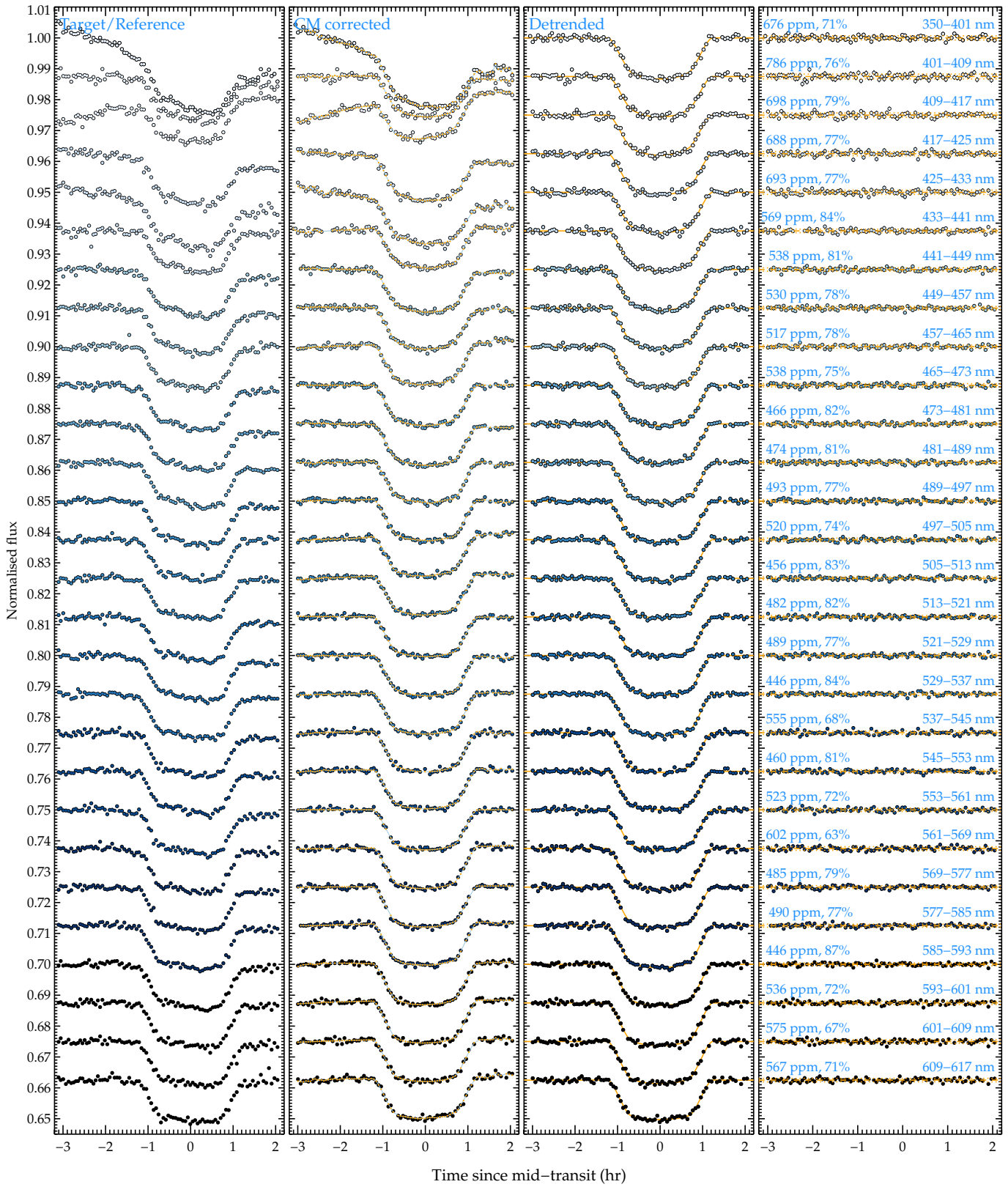
55. Rackham, B. et al. ACCESS I: an optical transmission spectrum of GJ 1214b reveals a heterogeneous stellar photosphere. *Astrophys. J.* **834**, 151 (2017).
56. Huitson, C. M. Gemini/GMOS transmission spectral survey: complete optical transmission spectrum of the hot Jupiter WASP-4b. *Astron. J.* **154**, 95–113 (2017).
57. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automatic Control* **19**, 716–723 (1974).
58. Gibson, N. P. Reliable inference of exoplanet light-curve parameters using deterministic and stochastic systematics models. *Mon. Not. R. Astron. Soc.* **445**, 3401–3414 (2014).
59. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
60. Pont, F. et al. The effect of red noise on planetary transit detection. *Mon. Not. R. Astron. Soc.* **373**, 231–242 (2006).
61. Lodders, K. Alkali element chemistry in cool dwarf atmospheres. *Astrophys. J.* **519**, 793–801 (1999).
62. Lodders, K. & Fegley, B. Atmospheric chemistry in giant planets, brown dwarfs, and low-mass dwarf stars. I. Carbon, nitrogen, and oxygen. *Icarus* **155**, 393–424 (2002).
63. Freedman, R. S. et al. Line and mean opacities for ultracool dwarfs and extrasolar planets. *Astrophys. J. Suppl. Ser.* **174**, 504–513 (2008).
64. Madhusudhan, N. & Seager, S. A temperature and abundance retrieval method for exoplanet atmospheres. *Astrophys. J.* **707**, 24–39 (2009).
65. Amundsen, D. S. et al. Accuracy tests of radiation schemes used in hot Jupiter global circulation models. *Astron. Astrophys.* **564**, A59 (2014).
66. Amundsen, D. S. et al. Treatment of overlapping gaseous absorption with the correlated-*k* method in hot Jupiter and brown dwarf atmosphere models. *Astron. Astrophys.* **598**, A97 (2017).
67. Tremblin, P. et al. Advection of potential temperature in the atmosphere of irradiated exoplanets: a robust mechanism to explain radius inflation. *Astrophys. J.* **841**, 30 (2017).
68. Drummond, B. The effects of consistent chemical kinetics calculations on the pressure-temperature profiles and emission spectra of hot Jupiters. *Astron. Astrophys.* **594**, A69 (2016).
69. Goyal, J. M. A library of ATMO forward model transmission spectra for hot Jupiter exoplanets. *Mon. Not. R. Astron. Soc.* **474**, 5158–5185 (2017).
70. Lecavelier des Etangs, A. et al. Rayleigh scattering in the transit spectrum of HD 189733b. *Astron. Astrophys.* **481**, L83–L86 (2008).
71. Eastman, J., Gaudi, B. S. & Agol, E. EXOFAST: a fast-exoplanetary fitting suite in IDL. *Publ. Astron. Soc. Pacif.* **125**, 83–112 (2013).
72. Drummond, B. et al. Observable signatures of wind-driven chemistry with a fully consistent three-dimensional radiative hydrodynamics model of HD 209458b. *Astrophys. J.* **855**, 31 (2018).
73. Wakeford, H. et al. The complete transmission spectrum of WASP-39b with a precise water constraint. *Astron. J.* **155**, 29–43 (2018).
74. Thorngren, D. P. et al. The mass-metallicity relation for giant planets. *Astrophys. J.* **831**, 64 (2016).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



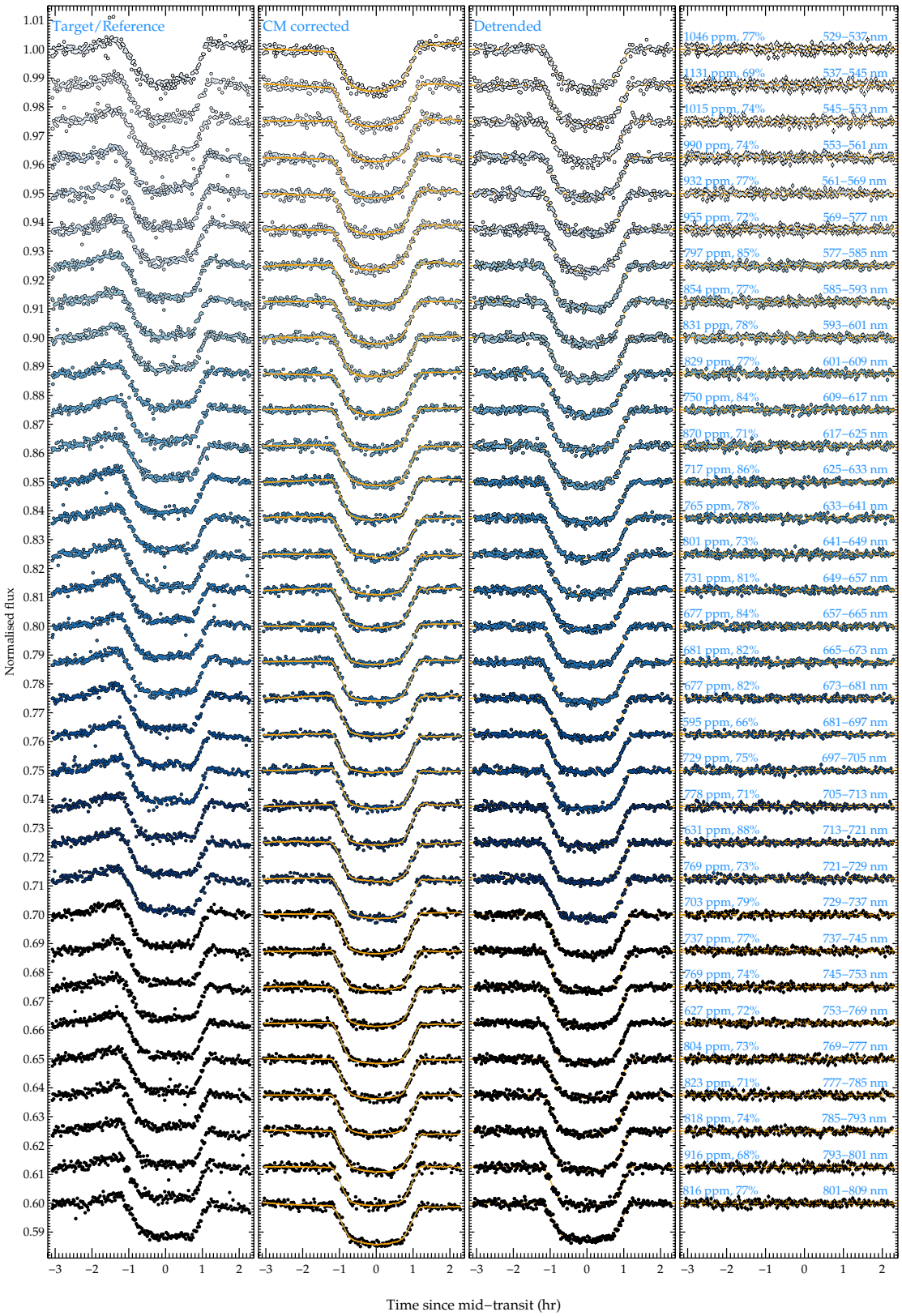
Extended Data Fig. 1 | VLT FORS2 stellar spectra and white-light curves. Left and right panels show the GRIS600B (blue) and GRIS600RI (red) datasets, respectively. The top row shows example stellar spectra used for relative spectrophotometric calibration. The dashed lines indicate the wavelength region used to produce the white-light curves. The second row shows normalized raw light curves for both sources. The third row shows normalized relative target-to-reference raw flux along with the marginalized Gaussian process model (A), the detrended transit light curve and model (B), and the common-mode correction (A/B). The fourth row shows the best-fit light curve residuals and 1σ error bars, obtained by subtracting the marginalized transit and systematics models from the relative target-to-reference raw flux. The two light curve residuals show dispersions of 78 and 201 parts per million, respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



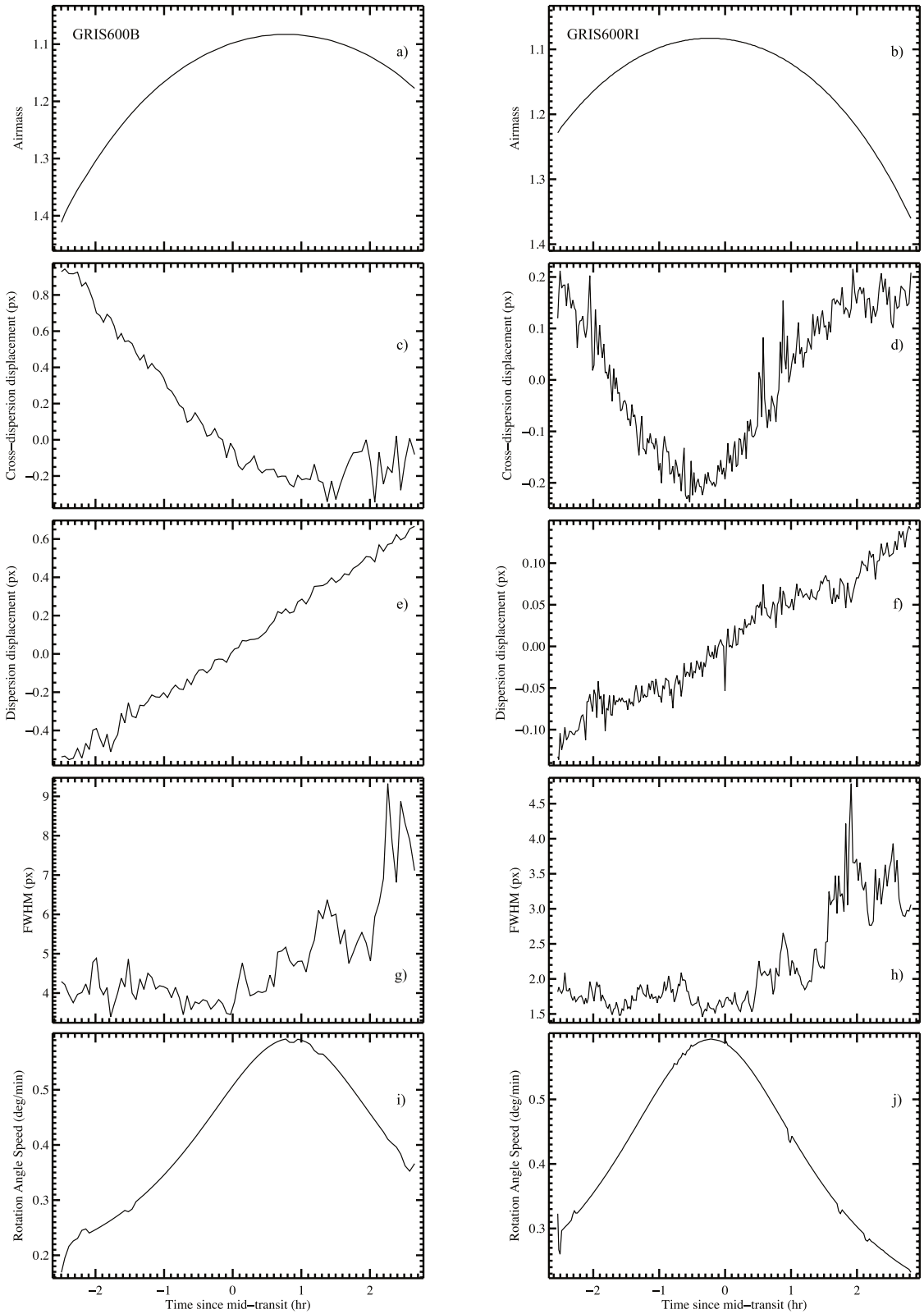
Extended Data Fig. 2 | Spectrophotometric light curves from grism 600B offset by a constant amount for clarity. The first panel shows the raw target-to-reference flux. The second panel shows the common-mode (CM)-corrected light curves and the transit and systematics models, with the highest statistical weight. The third panel shows detrended light curves and the transit model with the highest statistical weight. The fourth panel shows residuals with 1σ error bars. The dashed lines indicate the median residual level, with dotted lines indicating the dispersion and the percentage of the theoretical photon noise limit reached (blue).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



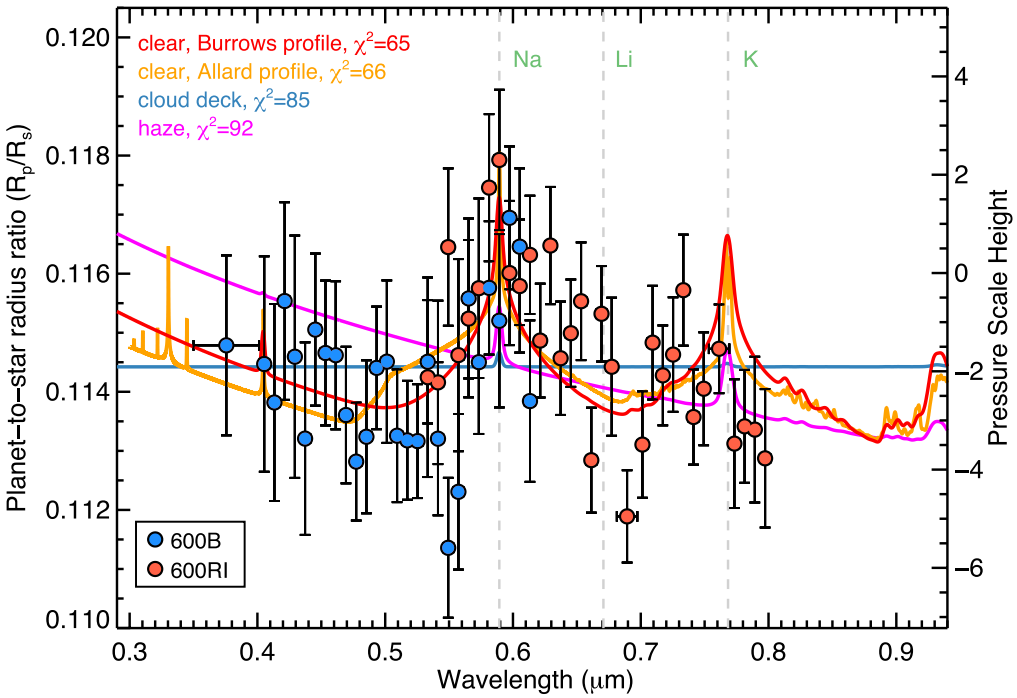
Extended Data Fig. 3 | As for Extended Data Fig. 2 but for grism 600RI.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



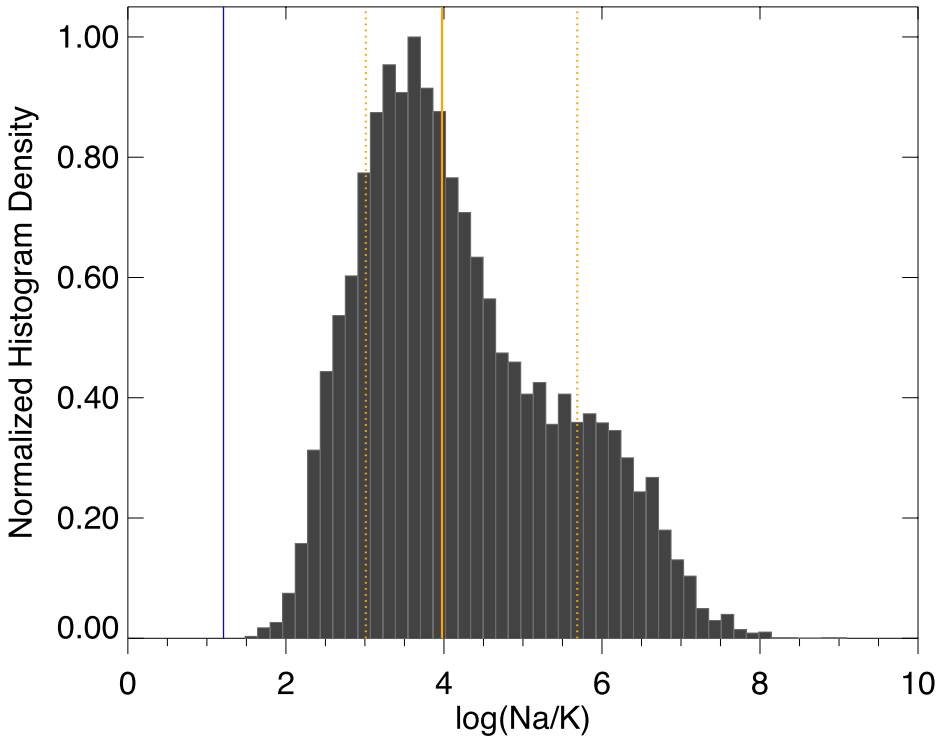
Extended Data Fig. 4 | Light-curve auxiliary variables. Shown are air mass (**a**, **b**), drifts along the cross-dispersion (**c**, **d**) and dispersion axes (**e**, **f**), FWHM (**g**, **h**) and the rate of change of the rotation angle (**i**, **j**) of the VLT FORS2 observations. Left and right columns refer to the GRIS600B and GRIS600RI observations, respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



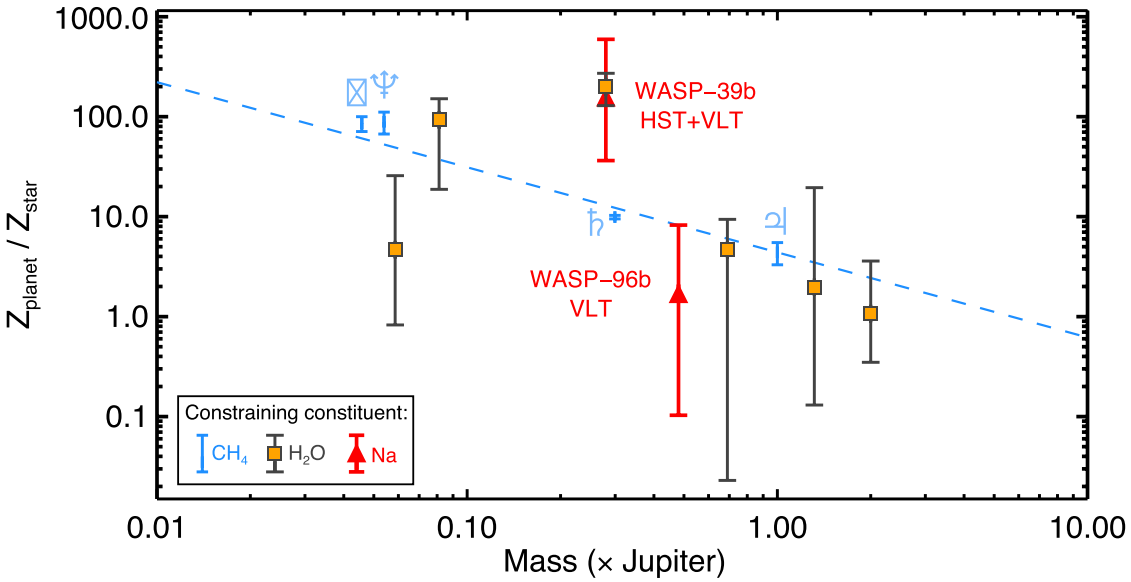
Extended Data Fig. 5 | Transmission spectrum of WASP-96b. Indicated are the relative radius measurements from grism 600B (blue dots) and 600RI (red dots) along with the 1σ uncertainties, compared to the same set of models as in Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



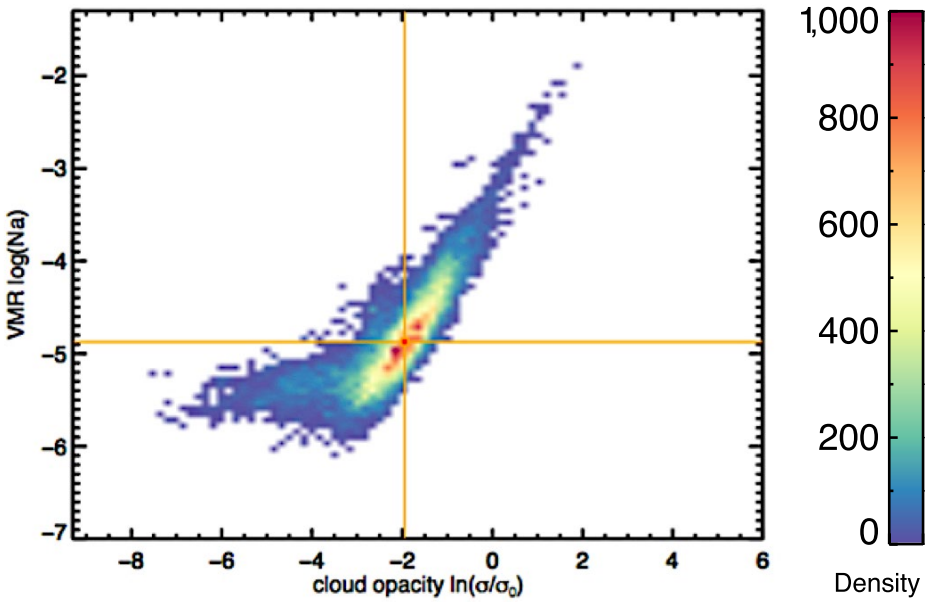
Extended Data Fig. 6 | Na/K ratio. Histogram of the marginalized posterior distribution of the Na to K ratio for WASP-96b. Shown are the median and 1σ levels (orange continuous and dotted lines, respectively). The solar value is indicated by the blue continuous line.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | Heavy element enrichment of exoplanets relative to their stars as a function of mass. Plotted are the Solar System planets (blue bars) and gas-giant exoplanets (grey and orange symbols). Each error bar represents the 1σ uncertainty. The blue line indicates a fit to the Solar System gas giants (pale blue symbols indicate Solar System planets).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 8 | Posterior distribution of the retrieved cloud opacity versus the sodium abundance. VMR, vertical mixing ratio. The median and 1σ measured parameters are indicated with continuous lines and the red dot marks the intercept for clarity. The colour scale shows the normalized density of the samples in the MCMC run.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | System parameters

Parameter	Value
Orbital period (day)	3.4252602 (fixed)
eccentricity	0 (fixed)
GRIS600B	4013 – 6173 Å
T_{mid} (JD)	$57963.33672^{+0.00034}_{-0.00033}$
i (°)	$85.21^{+0.18}_{-0.17}$
a/R_*	$8.93^{+0.18}_{-0.18}$
R_p/R_*	$0.1141^{+0.0015}_{-0.0014}$
u_1	$0.399^{+0.056}_{-0.030}$
u_2	$0.148^{+0.075}_{-0.040}$
A (ppm)	501^{+107}_{-72}
η_{FWHM} (arbitrary normalisation)	$1.01^{+1.2}_{-0.99}$
η_X (arbitrary normalisation)	$-3.45^{+0.70}_{-0.57}$
η_Y (arbitrary normalisation)	$-1.86^{+0.88}_{-1.01}$
c_0	$0.9992^{+0.00017}_{-0.00016}$
c_1	$-0.00068^{+0.00012}_{-0.00012}$
GRIS600RI	5268 – 8308 Å
T_{mid} (JD)	$57987.31195^{+0.00029}_{-0.00029}$
i (°)	$85.11^{+0.12}_{-0.12}$
a/R_*	$8.80^{+0.11}_{-0.12}$
R_p/R_*	$0.1172^{+0.0017}_{-0.0017}$
u_1	$0.26^{+0.11}_{-0.08}$
u_2	$0.23^{+0.12}_{-0.09}$
A (ppm)	2301^{+1690}_{-741}
η_{RAS} (arbitrary normalisation)	$0.50^{+0.87}_{-0.69}$
η_X (arbitrary normalisation)	$2.81^{+0.94}_{-0.76}$
η_Y (arbitrary normalisation)	$8.08^{+3.5}_{-2.5}$
c_0	$0.9980^{+0.0016}_{-0.0013}$
c_1	$0.00018^{+0.00029}_{-0.00028}$
Weighted mean:	
i (°)	85.14 ± 0.10
a/R_*	$8.84^{+0.09}_{-0.10}$
GRIS600B	(fixed i, a/R_* and T_{mid})
R_p/R_*	$0.1147^{+0.0014}_{-0.0014}$
u_1	$0.435^{+0.035}_{-0.020}$
u_2	$0.173^{+0.041}_{-0.022}$
GRIS600RI	(fixed i, a/R_* and T_{mid})
R_p/R_*	$0.1168^{+0.0015}_{-0.0014}$
u_1	$0.282^{+0.053}_{-0.032}$
u_2	$0.250^{+0.053}_{-0.032}$

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Transmission spectrum

λ (Å)	R_p/R_*	u_1	u_2
3500 – 4013	0.11479 ± 0.00152	0.484 ± 0.059	0.207
4013 – 4093	0.11447 ± 0.00182	0.553 ± 0.059	0.219
4093 – 4173	0.11382 ± 0.00167	0.539 ± 0.053	0.223
4173 – 4253	0.11554 ± 0.00167	0.493 ± 0.052	0.227
4253 – 4333	0.11459 ± 0.00205	0.458 ± 0.054	0.231
4333 – 4413	0.11321 ± 0.00163	0.464 ± 0.049	0.235
4413 – 4493	0.11505 ± 0.00129	0.479 ± 0.043	0.239
4493 – 4573	0.11466 ± 0.00123	0.522 ± 0.039	0.242
4573 – 4653	0.11462 ± 0.00125	0.415 ± 0.043	0.246
4653 – 4733	0.11361 ± 0.00116	0.422 ± 0.042	0.251
4733 – 4813	0.11282 ± 0.00100	0.404 ± 0.038	0.255
4813 – 4893	0.11324 ± 0.00130	0.385 ± 0.040	0.256
4893 – 4973	0.11440 ± 0.00104	0.337 ± 0.041	0.258
4973 – 5053	0.11451 ± 0.00138	0.367 ± 0.044	0.262
5053 – 5133	0.11326 ± 0.00112	0.410 ± 0.038	0.269
5133 – 5213	0.11318 ± 0.00101	0.288 ± 0.043	0.273
5213 – 5293	0.11317 ± 0.00096	0.336 ± 0.040	0.275
5293 – 5373	0.11444 ± 0.00089	0.260 ± 0.039	0.277
5373 – 5453	0.11365 ± 0.00095	0.349 ± 0.047	0.278
5453 – 5533	0.11361 ± 0.00089	0.299 ± 0.042	0.283
5533 – 5613	0.11322 ± 0.00102	0.330 ± 0.046	0.287
5613 – 5693	0.11541 ± 0.00095	0.303 ± 0.050	0.290
5693 – 5773	0.11499 ± 0.00095	0.284 ± 0.044	0.291
5773 – 5853	0.11653 ± 0.00084	0.263 ± 0.044	0.295
5853 – 5933	0.11685 ± 0.00092	0.270 ± 0.041	0.296
5933 – 6013	0.11648 ± 0.00086	0.236 ± 0.049	0.296
6013 – 6093	0.11607 ± 0.00086	0.242 ± 0.050	0.297
6093 – 6173	0.11545 ± 0.00081	0.169 ± 0.053	0.302
6173 – 6253	0.11487 ± 0.00096	0.316 ± 0.044	0.313
6253 – 6333	0.11648 ± 0.00099	0.235 ± 0.041	0.314
6333 – 6413	0.11457 ± 0.00096	0.194 ± 0.045	0.314
6413 – 6493	0.11499 ± 0.00091	0.185 ± 0.046	0.314
6493 – 6573	0.11553 ± 0.00100	0.267 ± 0.043	0.313
6573 – 6653	0.11284 ± 0.00089	0.250 ± 0.040	0.313
6653 – 6733	0.11532 ± 0.00081	0.192 ± 0.038	0.313
6733 – 6813	0.11442 ± 0.00117	0.155 ± 0.044	0.314
6813 – 6973	0.11189 ± 0.00078	0.087 ± 0.042	0.314
6973 – 7053	0.11311 ± 0.00090	0.100 ± 0.052	0.314
7053 – 7133	0.11483 ± 0.00096	0.129 ± 0.051	0.314
7133 – 7213	0.11428 ± 0.00085	0.104 ± 0.044	0.315
7213 – 7293	0.11463 ± 0.00097	0.156 ± 0.048	0.315
7293 – 7373	0.11572 ± 0.00094	0.277 ± 0.038	0.315
7373 – 7453	0.11357 ± 0.00081	0.102 ± 0.050	0.316
7453 – 7533	0.11405 ± 0.00096	0.263 ± 0.046	0.316
7533 – 7693	0.11473 ± 0.00075	0.147 ± 0.042	0.315
7693 – 7773	0.11312 ± 0.00109	0.190 ± 0.049	0.315
7773 – 7853	0.11341 ± 0.00096	0.189 ± 0.049	0.315
7853 – 7933	0.11336 ± 0.00123	0.175 ± 0.053	0.314
7933 – 8013	0.11287 ± 0.00117	0.214 ± 0.059	0.314

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0136-9

Fundamental limits to graphene plasmonics

G. X. Ni[1,2], A. S. McLeod[1,2], Z. Sun[2], L. Wang[3], L. Xiong[1,2], K. W. Post[2], S. S. Sunku[1,4], B.-Y. Jiang[2], J. Hone[3], C. R. Dean[1], M. M. Fogler[2] & D. N. Basov[1,2]*

Plasmon polaritons are hybrid excitations of light and mobile electrons that can confine the energy of long-wavelength radiation at the nanoscale. Plasmon polaritons may enable many enigmatic quantum effects, including lasing[1], topological protection[2,3] and dipole-forbidden absorption[4]. A necessary condition for realizing such phenomena is a long plasmonic lifetime, which is notoriously difficult to achieve for highly confined modes[5]. Plasmon polaritons in graphene—hybrids of Dirac quasiparticles and infrared photons—provide a platform for exploring light–matter interaction at the nanoscale[6,7]. However, plasmonic dissipation in graphene is substantial[8] and its fundamental limits remain undetermined. Here we use nanometre-scale infrared imaging to investigate propagating plasmon polaritons in high-mobility encapsulated graphene at cryogenic temperatures. In this regime, the propagation of plasmon polaritons is primarily restricted by the dielectric losses of the encapsulated layers, with a minor contribution from electron–phonon interactions. At liquid-nitrogen temperatures, the intrinsic plasmonic propagation length can exceed 10 micrometres, or 50 plasmonic wavelengths, thus setting a record for highly confined and tunable polariton modes. Our nanoscale imaging results reveal the physics of plasmonic dissipation and will be instrumental in mitigating such losses in heterostructure engineering applications.

Here we investigate plasmon polariton propagation and dissipation using near-field infrared microscopy, enabling a direct visualization of polaritonic standing waves on the surface of graphene and other van der Waals materials[6,7,9,10]. Both the wavelength of plasmon polariton waves, λ_p, and their travel range, which quantifies their dissipation, can be readily confirmed from nanoscale infrared images (Figs. 1, 2). These two observables permit direct extraction of the response functions of a plasmonic medium, including the complex conductivity $\sigma(\omega)=\sigma'+i\sigma''$, as we detail below. The response functions encode information on both interactions and scattering processes in an electron liquid. Thus, near-field examination of plasmon polariton images facilitates inquiry into fundamental electronic phenomena of the media supporting these polaritons. As customary in solid-state physics, scattering processes of diverse origin can be distinguished by investigating their dependence on control parameters—in particular, temperature. This motivates a plasmonic imaging investigation of graphene at cryogenic temperatures with the help of a newly developed apparatus[11]. We examined a high-mobility micro-device constructed from a graphene monolayer encapsulated between two thin slabs of hexagonal boron nitride (hBN), with a graphene carrier density that is tunable through a silicon back gate (Fig. 1a, b). The encapsulation preserved the graphene in its pristine form[12]. Our cryo-imaging approach to recording plasmon polariton standing waves, together with a theoretical model free of adjustable parameters, has enabled us to identify the physics governing plasmon propagation in high-mobility graphene.

Nanoscale imaging of plasmon polaritons is achieved using the following experimental protocol. The metallized tip of an atomic force microscope (AFM) probe is illuminated by a focused laser field with an infrared frequency of ω. The tip functions as an optical antenna that intensifies the incident electric field at the apex. When the tip is brought near the sample, this concentrated field can excite polaritonic modes with a wavelength of $\lambda_p(\omega)$ that is much shorter than the free-space wavelength $\lambda_{IR}=2\pi/\omega$ of the incident light[13]. The optical antenna functionality of the AFM tip also facilitates scattering of local near fields back into the far field, where they are registered by conventional optical detection. Proper demodulation methods allow one to isolate the back-scattered light component that is associated with the local electric field confined to the area under the tip, thus achieving a spatial resolution of the order of 10 nm (ref. [14]). In such a near-field imaging experiment, plasmon polaritons manifest as a periodic modulation (fringes) of the observed near-field signal as a function of the tip position. Similarly to imaging experiments at room temperature[8], fringes of two distinct periodicities, λ_p and $\lambda_p/2$, appear in Figs. 1 and 2. The $\lambda_p/2$-period fringes are due to plasmon polaritons that are emitted by the tip and complete a round-trip between the tip and reflection from the sample edges. The λ_p-period fringes are produced by the interference between tip-launched plasmon polaritons and the evanescent component of the reflected ones[15]. Alternatively, the λ_p-period fringes

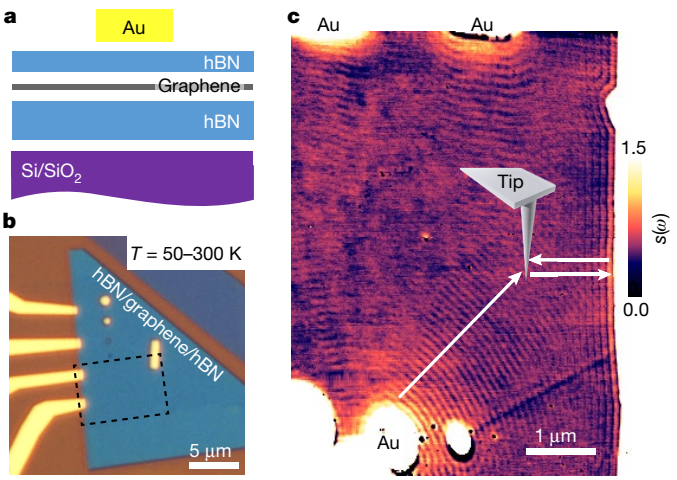


Fig. 1 | Nanoscale infrared imaging of surface plasmons in Au/hBN/graphene/hBN encapsulated structures at cryogenic temperature. **a**, Sketch of the layered Au/hBN/graphene/hBN/SiO_2/Si heterostructure. The lithographically defined gold (Au) microstructures on top of the hBN capping layer serve as effective plasmonic launchers. **b**, Optical image of the device. The black dashed rectangle marks the area shown in **c**. The entire field of view is within the diffraction-limited spot of our infrared laser, which operates at an energy of 886 cm^{-1} (110 meV) or $\lambda_{IR}=11.28\,\mu m$. **c**, Nanoscale infrared image of plasmonic interference fringes from the encapsulated graphene monolayer, expressed by the normalized scattering amplitude s acquired at a back-gate voltage of 97 V and a temperature of $T=60$ K. The arrows represent the propagation direction of the plasmon waves. These experiments simultaneously visualize the local electric field associated with interference from plasmon polaritons emitted by the near-field probe and reflected by sample edges, as well as from polaritons emitted by the gold microstructures (labelled as Au).

[1]Department of Physics, Columbia University, New York, NY, USA. [2]Department of Physics, University of California, San Diego, La Jolla, CA, USA. [3]Department of Mechanical Engineering, Columbia University, New York, NY, USA. [4]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA. [5]These authors contributed equally: G. X. Ni, A. S. McLeod. *e-mail: db3056@columbia.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

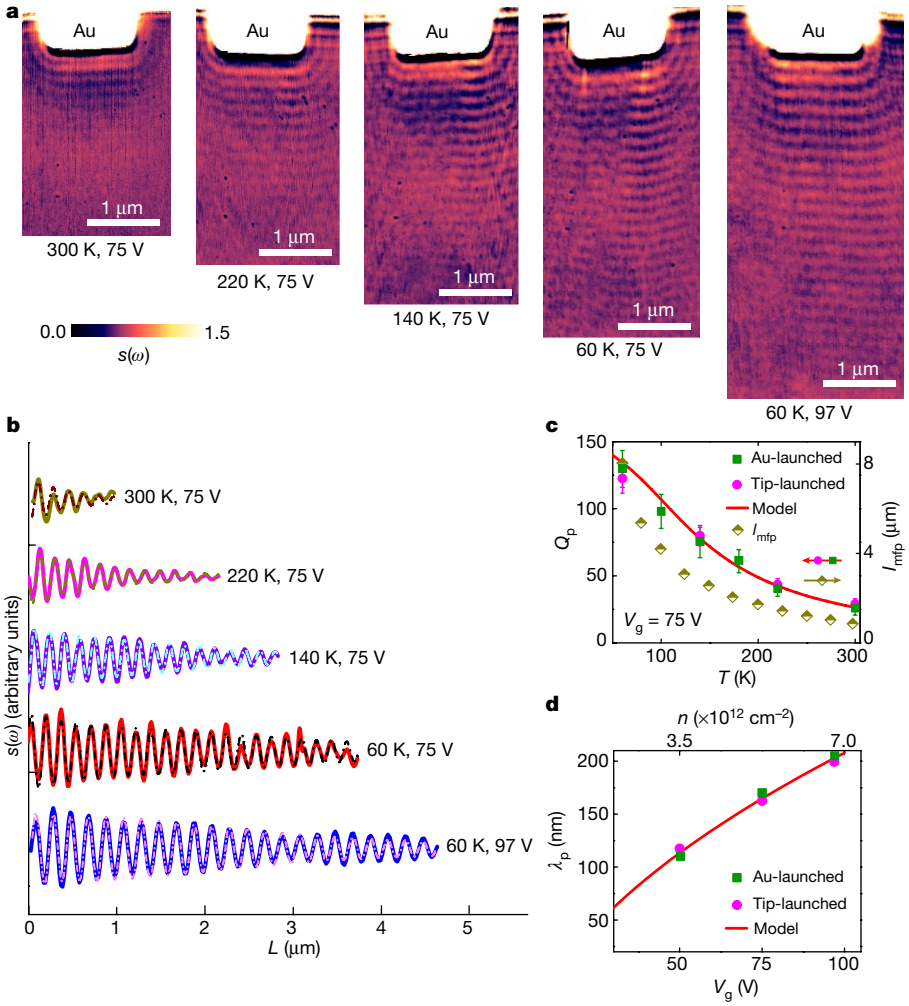


Fig. 2 | Temperature- and gate-dependent trends in surface plasmon propagation in graphene. a, Nanoscale infrared images of the normalized scattering amplitude s acquired at sequential sample temperatures and gate voltages. A gold electrode (labelled Au at the top of the images) functions as an antenna that emits graphene plasmons. **b**, Line profiles of plasmonic interference fringes propagating from left to right, as a function of the distance L from the gold launcher. The variable attenuation of the propagating waveform (solid curves) is characteristic of the temperature-dependent plasmon scattering rate. Dash-dotted lines represent the results of numerical simulations performed to identify the temperature dependence of the complex plasmon wavevector and associated scattering rate (see main text and Methods). **c**, The temperature dependence of the quality factor Q_p for both λ_p- and $\lambda_p/2$-period plasmon waves obtained from nanoscale infrared images taken at $V_g = 75$ V. The red solid curve is the theoretical model detailed in Methods. The electronic mean free path $l_{mfp}(T)$ is also plotted for comparison. The error bars define the 95% confidence intervals. **d**, The voltage (V_g) and carrier density (n) dependence of the plasmon wavelength, obtained from nanoscale infrared data for both λ_p- and $\lambda_p/2$-period plasmonic patterns. The red solid curve shows a $\lambda_p \propto \sqrt{V_g}$ dependence, in accord with previous results[9,17].

can arise from plasmon polaritons launched by emitters other than the tip that are present in the device[8]. The efficiency of the plasmon polariton emission depends on the dielectric polarizability, shape and size of the emitter. Specifically, graphene edges are weak emitters, but elongated metallic objects placed on graphene can act as efficient ones[16]. In our structures, the gold contacts on top of the hBN-encapsulating layer act as fixed plasmonic antennas (Figs. 1, 2).

We begin with a survey of a large-area (6 μm × 8 μm) image of plasmon polariton standing waves obtained at $T = 60$ K (Fig. 1c) with an infrared laser operating at $\lambda_{IR} = 11.28$ μm and at a back-gate voltage of $V_g = 97$ V. Here we display raw data in the form of the amplitude s of the scattered field, normalized by the amplitude detected at the gold contacts, whose optical response provides a convenient reference that is independent from temperature and carrier density (see Methods for details). The most prominent aspect of the image in Fig. 1c is that the entire field of view is filled with plasmonic interference fringes. The $\lambda_p/2$-period fringes are most evident on the right-hand side of the field of view, close to a natural boundary of the graphene monolayer, whereas λ_p-period fringes dominate in the vicinity of gold antennas. Figure 1c reveals that plasmon polaritons are highly confined, with $\lambda_{IR}/\lambda_p > 60$; however, they travel over several micrometres, far exceeding previous benchmark results at room temperature[8,15].

The nanoscale infrared images and corresponding line profiles in Fig. 2 attest to a dramatic reduction of plasmonic losses as the temperature of the graphene is reduced. Figure 2a shows detailed raster scans obtained in the vicinity of one of the gold antennas at $V_g = 75$ V and at temperatures between 60 K and 300 K. When the gate voltage increases above $V_g = 50$ V, plasmonic fringes emerge out of a featureless background, as illustrated in a voltage-dependent scan shown in Extended Data Fig. 3. At room temperature, plasmon polariton propagation is restricted to distances of less than 1 μm. As the temperature is lowered, we observe a systematic increase of both the overall travel range and the number of detectable fringes. At $T = 60$ K plasmonic oscillations extend beyond 4 μm at $V_g = 75$ V and beyond 5 μm at $V_g = 97$ V (Fig. 2a, b). These overall trends are in accord with the temperature dependence of the electronic mean free path, l_{mfp}, in high-mobility devices prepared under identical protocols[12]. In fact, even the magnitudes of the plasmon polariton travel range and of l_{mfp} appear to be on par with each other: l_{mfp} increases from 1 μm at $T = 300$ K up to 8 μm at 60 K (Fig. 2c). The latter value exceeds the entire field of view in Fig. 2a and is consistent with the notion of ballistic electronic transport. The novelty of the images in Fig. 2 is that they manifest the real space behaviour of graphene plasmon polaritons supported by ballistic electrons.

To quantify the quasiparticle dynamics underlying the propagation of plasmon polaritons in graphene, we present an analysis based on the complex momentum of the polariton, $q_p = q_p' + iq_p''$. The real part, q_p', defines the plasmon polariton wavelength through $\lambda_p = 2\pi/q_p'$, whereas the imaginary part, q_p'', quantifies dissipation and the quality factor $Q_p = q_p'/q_p''$ (Table 1). The complex momentum of plasmon polaritons, and thus both λ_p and Q_p, are governed by the optical conductivity of graphene (see Methods section 'Plasmons near a graphene edge') or, alternatively, can be obtained through numerical analysis of plasmonic fringes (Methods section 'Plasmonic damping rate analysis'). The evolution of λ_p with the gate voltage (or carrier density) is presented in Fig. 2d and is in complete accord with the results of earlier studies at room temperature[8–10,15,17]. A noteworthy aspect of our data is the rapid growth of the quality factor at low temperatures (Fig. 2c).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Table 1 | Comparison of graphene plasmon parameters with those of metals (Ag) and doped semiconductors (n-doped InSb and CdO)

	Confinement ratio	Quality factor	Lifetime (fs)
Definition	λ_{IR}/λ_p	$Q_p = q_p'/q_p''$	$\tau = 2Q_p/\omega$
Graphene (experimental, $T = 60$ K)	66	130	1,600
Graphene (intrinsic, $T = 60$ K)	66	970	12,000
Ag[a,b] ($T = 10$ K)	~1	36	14
n-InSb[c], n-CdO[a] ($T = 300$ K)	<10	37	270

Parameters relevant for graphene entries: frequency, $\omega/(2\pi c) = 886\,\mathrm{cm}^{-1}$; gate voltage, $V_g = 75$ V; plasmon wavelength, $\lambda_p = 170$ nm; and dielectric loss of the substrate, $\gamma_z = 5.9\,\mathrm{cm}^{-1}$ (see Methods for details).
[a]Data from ref. [29].
[b]Data from ref. [30].
[c]Data from ref. [31].

This temperature dependence tracks the variation of the electronic mean free path. We therefore surmise that the evolution of $Q_p(T)$ reflects the same fundamental physics as that governing electronic transport in graphene, but probed at finite frequency ω. The highest quality factor detected in our plasmonic imaging experiments is $Q_p = 130$.

A detailed quantitative analysis of finite-frequency dissipation of Dirac quasiparticles requires that intrinsic dissipation in graphene is distinguished from extrinsic losses associated with the dielectric environment of graphene. We outline this procedure in Methods section 'Plasmonic damping rate analysis'. In Fig. 3a we show the plasmonic scattering rate $\gamma_p(T) = \omega\sigma'/\sigma''$, which quantifies intrinsic dissipation originating from within the graphene monolayer. Inspection of these data reveals a dramatic drop in the intrinsic plasmonic scattering rate from about 20 cm^{-1} at room temperature to $<2.0\,\mathrm{cm}^{-1}$ at $T = 60$ K (we note the frequency unit conversion rule: $1\,\mathrm{cm}^{-1} = 30\,\mathrm{GHz} = 1.43$ K). The corresponding direct current (d.c.) scattering rate measured for a similar encapsulated device is also plotted in Fig. 3a, showing the same overall trend in temperature dependence. Another important finding reported here is that the plasmonic scattering rate exceeds the transport scattering rate by a factor of two at room temperature. Notably, the same physical model (blue and red lines in Fig. 3a), which is free of adjustable parameters, yields a consistent description of the dissipation encountered by Dirac quasiparticles in both d.c. transport and at a finite frequency in the form of collective plasmon polaritons.

We now outline the key elements that contribute to d.c. and plasmonic scattering in our model. A substantial contribution to plasmon polariton losses is from acoustic phonons, as reported in refs [18,19]. It is practical to discriminate between two distinct loss channels associated with acoustic phonons, which in the related literature are commonly referred to as the deformation potential effect and the pseudo-magnetic field effect. The deformation potential effect shifts the energy of Dirac bands (Fig. 3d), whereas the pseudo-magnetic field effect displaces the overall momentum of Dirac cones in opposite directions for the two valleys of the Brillouin zone at K and K′ (see Fig. 3d and Methods section 'Plasmonic damping rate analysis'). First-principle calculations show that the deformation potential channel is vanishingly small[20,21], leaving the pseudo-magnetic field component as the dominant acoustic scatterer. However, the pseudo-magnetic field effect yields a conventional linear temperature dependence of the quasiparticle scattering rate that is nearly indistinguishable for d.c. scattering and plasmonic scattering at $\omega = 886\,\mathrm{cm}^{-1}$ (orange and blue lines in Fig. 3b). This finding prompted a search for scattering mechanisms with a pronounced frequency dependence. Such a contribution emerges from inter-valley electron–phonon scattering, with scattering rate γ_K (black and red lines in Fig. 3b). The matrix element used to numerically evaluate γ_K requires the knowledge of both the energy of the zone-boundary optical phonon of graphene, ω_K, and the electron–phonon coupling constant, β_K, for which we used the results of prior first-principles calculations[20] (Methods section 'Plasmonic damping rate analysis'). The physical mechanism behind the frequency dependence of γ_K is the thermally activated nature of phonon-mediated inter-valley scattering, which yields a scattering rate proportional to $\exp[-(\omega_K - \omega)/T]$, where $\omega_K = 1{,}200\,\mathrm{cm}^{-1}$. An additional small contribution to plasmon polariton damping comes from electron–electron scattering, with scattering rate γ_{ee} (magenta line in Fig. 3b). The net result is that three added channels—the pseudo-magnetic field effect, inter-valley scattering and electron–electron scattering—reproduce the observed temperature dependence of both d.c. and plasmonic scattering measured for ultrahigh-mobility encapsulated graphene. We stress that our scattering rate analysis involves accurate evaluation of the temperature-dependent losses in hBN microcrystals that are nominally identical to those used in the actual graphene structures (Methods section 'The optical constants of hBN'). This procedure enabled the accurate decomposition of various scattering mechanisms and thus enhanced the reliability of our findings.

Our experimental results, together with our unified parameter-free theoretical analysis, account for quasiparticle dynamics in ultrahigh-mobility graphene, even at finite frequencies ω relevant for graphene plasmonics. Our experimental values of the quality factor Q_p and the lifetime τ of plasmons in graphene, extracted from cryogenic imaging, are higher than those of any other materials used for infrared plasmonics (Table 1). Although limited by acoustic-phonon scattering, the intrinsic propagation length $l_p = 1/(2q_p'')$ of plasmon polaritons in our high-mobility structures can exceed 13 μm at low temperatures (the intrinsic quality factor is $Q_p > 970$; Table 1). In this regime, transport

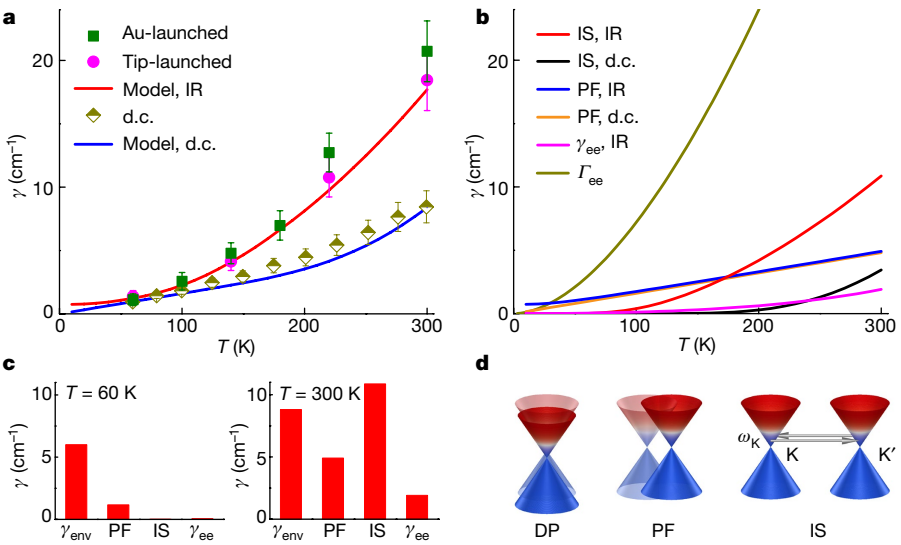


Fig. 3 | Plasmonic and electronic transport in high-mobility graphene. a, Temperature dependence of the plasmonic scattering rate (squares and circles) and of the d.c. scattering rate (diamonds). Solid lines display the results of our parameter-free modelling. The error bars represent 95% confidence intervals. **b**, Key contributions to plasmonic and electronic scattering and their temperature dependences. **c**, Histograms displaying the contribution of different scattering channels at $T = 60$ K (left) and 300 K (right). **d**, Schematics of transformations of graphene electronic structure associated with the deformation potential (DP), pseudo-magnetic field (PF) and inter-valley scattering (IS). The arrows in denote the possible scattering path across the K and K′ valleys.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

of plasmon polaritons is dominated by the dielectric environment and can be limited by the physical dimensions of macroscopic devices. Although losses associated with the encapsulating hBN layers and SiO_2 are extrinsic to graphene, dielectrics are practically unavoidable in actual high-mobility graphene devices. At ambient conditions, losses from (1) the dielectric layers, (2) acoustic-phonon scattering and (3) inter-valley scattering all sum together with unequal contributions. The dielectric losses can be reduced by 20% by using a thicker hBN bottom layer to increase the distance between the lossier SiO_2 and graphene. However, the other two mechanisms are more difficult to circumvent. We stress that the contribution of electron–electron scattering to finite-frequency plasmon damping is small, even though the electron–electron collision rate Γ_{ee} is comparable to the electron–phonon scattering rate (see Fig. 3b and Methods section 'Plasmonic damping rate analysis'). In general, electronic interactions can influence the conductivity only if the Galilean invariance of the system is broken, which is a weak effect in our high-carrier-density graphene at low temperature[22]. Our data imply that the contribution of the electron–electron interaction to plasmonic responses is more prominent at higher temperatures, at lower carrier densities or in samples of special geometry. However, a major change in the electrodynamic response should occur at frequencies below the electron–electron collision rate Γ_{ee}. In that regime (called the hydrodynamic regime), which has been recently studied in d.c. transport experiments[23,24], the plasmons are expected to become overdamped by another type of collective modes: energy waves or 'demon' modes[22,25]. These modes could be studied using terahertz near-field microscopy combined with photocurrent detection[26]. Finally, we note that the extrinsic plasmonic de-phasing is surprisingly weak in many macroscopic areas of Fig. 1c on which we focused our analysis, whereas its presence is obvious in other areas. These remaining extrinsic limitations on plasmon polariton propagation, including the possible role of sample boundaries[27], will be a subject of future nanoscale imaging studies. Nevertheless, this first report of plasmonic nanoscale imaging studies of encapsulated graphene at cryogenic temperatures demonstrates potential for the exploration of plasmonic switching and nonlinear phenomena in ultrahigh-mobility graphene at the liquid-nitrogen temperatures that are typically used for mid-infrared technologies. Besides imposing fundamental limits on plasmon damping, the electron–phonon interaction probed here may have a role in other many-body phenomena, such as the recently discovered superconductivity in twisted bilayer graphene[28].

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0136-9.

Received: 7 November 2017; Accepted: 26 March 2018;
Published online 23 May 2018.

1. Lu, Y. J. et al. Plasmonic nanolaser using epitaxially grown silver film. *Science* **337**, 450–453 (2012).
2. Jin, D. et al. Infrared topological plasmons in graphene. *Phys. Rev. Lett.* **118**, 245301 (2017); erratum **119**, 019901 (2017).
3. Karzig, T., Bardyn, C. E., Lindner, N. H. & Refael, G. Topological polaritons. *Phys. Rev. X* **5**, 031001 (2015).
4. Rivera, N., Kaminer, I., Zhen, B., Joannopoulos, J. D. & Soljacic, M. Shrinking light to allow forbidden transitions on the atomic scale. *Science* **353**, 263–269 (2016).
5. Boltasseva, A. & Atwater, H. A. Low-loss plasmonics metamaterials. *Science* **331**, 290–291 (2011).
6. Basov, D. N., Fogler, M. M. & Garcia de Abajo, F. J. Polaritons in van der Waals materials. *Science* **354**, aag1992 (2016).
7. Low, T. et al. Polaritons in layered two-dimensional materials. *Nat. Mater.* **16**, 182–194 (2017).
8. Woessner, A. et al. Highly confined low-loss plasmons in graphene–boron nitride heterostructures. *Nat. Mater.* **14**, 421–425 (2015).
9. Fei, Z. et al. Gate-tuning of graphene plasmons revealed by infrared nano-imaging. *Nature* **487**, 82–85 (2012).
10. Chen, J. et al. Optical nano-imaging of gate-tunable graphene plasmons. *Nature* **487**, 77–81 (2012).
11. McLeod, A. S. et al. Nanotextured phase coexistence in the correlated insulator V_2O_3. *Nat. Phys.* **13**, 80–86 (2017).
12. Wang, L. et al. One-dimensional electrical contact to a two-dimensional material. *Science* **342**, 614–617 (2013).
13. Novotny, L. & Hecht, B. *Principles of Nano-Optics* (Cambridge Univ. Press, Cambridge, 2006).
14. Atkin, J. H., Berweger, S., Jones, A. C. & Raschke, M. B. Nano-optical imaging and spectroscopy of order, phases, and domains in complex solids. *Adv. Phys.* **61**, 745–842 (2012).
15. Ni, G. X. et al. Ultrafast optical switching of infrared plasmon polaritons in high-mobility graphene. *Nat. Photon.* **10**, 244–247 (2016).
16. Alonso-Gonzalez, P. et al. Controlling graphene plasmons with resonant metal antennas and spatial conductivity patterns. *Science* **344**, 1369–1373 (2014).
17. Ni, G. X. et al. Plasmons in graphene moiré superlattices. *Nat. Mater.* **14**, 1217–1222 (2015).
18. Principi, A. et al. Plasmon losses due to electron-phonon scattering: the case of graphene encapsulated in hexagonal Boron Nitride. *Phys. Rev. B* **90**, 165408 (2014).
19. Hwang, E. & Das Sarma, S. Acoustic phonon scattering limited carrier mobility in two-dimensional extrinsic graphene. *Phys. Rev. B* **77**, 115449 (2008).
20. Park, C. H. et al. Electron-phonon interactions and the intrinsic electrical resistivity of graphene. *Nano Lett.* **14**, 1113–1119 (2014).
21. Sohier, T., Calandra, M. & Mauri, F. Density-functional calculation of static screening in two-dimensional materials: the long-wavelength dielectric function of graphene. *Phys. Rev. B* **91**, 165428 (2015).
22. Sun, Z., Basov, D. N. & Fogler, M. M. Universal linear and nonlinear electrodynamics of a Dirac fluid. *Proc. Natl Acad. Sci. USA* **115**, 3285–3289 (2018).
23. Bandurin, D. A. et al. Negative local resistance caused by viscous electron backflow in graphene. *Science* **351**, 1055–1058 (2016).
24. Kumar, R. K. et al. Superballistic flow of viscous electron fluid through graphene constrictions. *Nat. Phys.* **13**, 1182 (2017).
25. Phan, T. V., Song, J. C. W. & Levitov, L. S. Ballistic heat transfer and energy waves in an electron system. Preprint at https://arxiv.org/abs/1306.4972 (2013).
26. Lundeberg, M. et al. Tuning quantum non-local effects in graphene plasmonics. *Science* **357**, 187–191 (2017).
27. Halbertal, D. et al. Imaging resonant dissipation from individual atomic defects in graphene. *Science* **358**, 1303–1306 (2017).
28. Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
29. Caldwell, J. D. et al. Low-loss, infrared and terahertz nanophotonics using surface phonon polaritons. *Nanophotonics* **4**, 44–68 (2014).
30. Sun, L. et al. Enhancement of plamonic performance in epitaxial silver at low temperature. *Sci. Rep.* **7**, 8917 (2017).
31. Chochol, J. et al. Plasmonic behavior of III–V semiconductor in far-infrared and terahertz range. *J. Eur. Opt. Soc.-Rapid* **13**, 13 (2017).

Acknowledgements We thank A. Charnukha, A. Frenzel, R. Ribeiro-Palau and A. Sternbach for discussions. Research on Dirac quasiparticle dissipation in graphene was supported by DOE-BES DE-SC0018426. Plasmonic nanoscale imaging at cryogenic temperatures was supported by DOE-BES DE-SC0018218. Work on infrared nanoscale antennas and metasurfaces was supported by AFOSR FA9550-15-1-0478. The development of scanning plasmon interferometry was supported by ONR N00014-15-1-2671. Upgrades of the ultrahigh vacuum scanning probe system were supported by ARO grant W911nf-17-1-0543. D.N.B was supported by the Gordon and Betty Moore Foundation's EPiQS Initiative through Grant GBMF4533. J.H. acknowledges support from ONR N00014-13-1-0662.

Reviewer information *Nature* thanks J. Song, A. Zayats and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions G.X.N. and A.S.M. performed the nanoscale infrared measurements and characterizations. Z.S., L.X., B.-Y.J. and M.M.F. provided theoretical calculations. L.W., J.H. and C.R.D. designed and created the device structures. G.X.N., L.X., K.W.P. and S.S.S. performed far-field optical spectroscopy measurements and characterizations. D.N.B. supervised the project. G.X.N., Z.S., M.M.F. and D.N.B. co-wrote the manuscript with input from all co-authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0136-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.N.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Device fabrication and characterization. Our high-quality hBN/graphene/hBN devices were fabricated using a polymer-free multilevel stacking dry-transfer process to form the encapsulated graphene structure[12]. Briefly, a very thin (7 nm) hBN crystal was used to pick up single-layer graphene and another hBN crystal in sequence. Bubble- and wrinkle-free stacks were then transferred onto SiO_2/Si substrates. For the etching process, only a small portion of the graphene edge was exposed for edge-contact deposition (Extended Data Fig. 1). Thus, most of the graphene specimen remained intact. This technique is essential for maintaining graphene microcrystals free of contamination, bubbles and wrinkles. Consequently, the charge neutrality point of graphene in these samples is very close to zero gate voltage at ambient conditions, as confirmed from our near-field gating experiments (Extended Data Fig. 1). After multi-stacking transfer, metal contacts were fabricated at the exposed graphene edge. We note that graphene was entirely encapsulated within the hBN flakes.

Electrical contacts to graphene were provided by a one-dimensional metallic edge of deposited gold, permitting completely polymer-free assembly of the device layers[12]. Together with a silicon back gate fabricated under 285 nm of thermal oxide, these contacts enabled in situ tuning of the graphene carrier density to values as high as 7×10^{12} cm^{-2} for electrons and 5×10^{12} cm^{-2} for holes during the imaging experiment. During our nanoscale infrared imaging experiments, these metallic contacts were also used as effective emitters of graphene plasmons through their efficient in-coupling (scattering) of incident infrared light into polaritons, as reported previously[16] and as detailed below. Our device includes a thin, 7-nm-thick top layer of hBN to facilitate nanoscale infrared imaging of the embedded graphene monolayer owing to the finite penetration depth of electric near-fields and subsurface detection capabilities of nanoscale infrared imaging[8,15]. The hBN is optically transparent at our imaging energy of 886 cm^{-1} (110 meV), ensuring that our nanoscale infrared i plasmons show oscillations with a single p mages strictly document the plasmonic optical response of graphene.

The AFM topography image of the device is shown in Extended Data Fig. 1a. Extended Data Fig. 1b and c shows room-temperature near-field images obtained at gate voltages $V_g = 0$ V and 50 V, respectively. In Extended Data Fig. 1b, the near-field shows practically no contrast resulting from the presence of graphene crystal. This confirms that the encapsulation by hBN kept graphene nearly charge-neutral at ambient conditions. By contrast, Extended Data Fig. 1c clearly shows a rectangular graphene shape due to infrared contrast arising from the overall conductivity enhancement under gating. Moreover, multiple plasmonic fringes along the Au antenna and graphene edges can be detected, which are discussed in the main text and also in 'Plasmons near a graphene edge' below.

Cryogenic infrared nanoscale imaging experiments. The nanoscale infrared imaging experiments were performed using a home-built scattering-type scanning near-field optical microscope (SNOM) operating with variable sample temperatures of 20–450 K. All measurements were conducted under ultrahigh vacuum conditions at a pressure lower than 10^{-9} mbar to prevent sample-surface contamination. Further technical details of the microscope are documented in ref. [11]. The cryogenic scattering-type SNOM is equipped with continuous-wave mid-infrared quantum cascade lasers (DRS Daylight Solutions) and CO_2 lasers (Access Laser). The scattering-type SNOM is based on an AFM, which in the present experiments operated in non-contact mode using cantilevered metallic AFM probes with a tip apex radius of ~25 nm and tapping frequencies of ~270 kHz. A pseudo-heterodyne interferometric detection module is implemented to extract both the scattering amplitude s and the phase of the near-field signal; here, we discuss only the former. To suppress background contributions to the back-scattered near-field signal, we demodulated the detected signal at the third harmonic of the probe tapping frequency.

Plasmons near a graphene edge. In the main text, we focused on plasmons launched by the Au pad. In this section, we present temperature-dependent data for plasmons launched by the tip of the scattering-type SNOM near the graphene edges. Extended Data Fig. 2a shows raster-scanned images of the edge regions at a gate voltage of $V_g = 75$ V, taken over a temperature range of $T = 60$–300 K. At $T = 300$ K, plasmons propagate to about 1 μm from the edge of the sample, consistent with the earlier experiments[8,15]. As the temperature is lowered, both the range and the number of detectable plasmonic fringes increase. At $T = 60$ K, the number of plasmonic oscillations almost doubles compared to the room temperature. This behaviour is consistent with what is observed for plasmons launched by the Au pad (Fig. 2).

Whereas images of the Au-launched plasmons show oscillations with a single period, λ_p, the tip-launched plasmons near the edge exhibit two distinct periodicities, λ_p and $\lambda_p/2$, as a function of the tip–edge distance L'. According to previous studies[8,15], the $\lambda_p/2$ period is due to the standing wave produced by plasmons making a round-trip from the tip to the edge and back. In addition to the standing wave, the total electric field in the system contains a component that does not oscillate, but instead decays as a power-law of the distance x from the edge. This part of the total field, which exists because of the long-range nature of the Coulomb interaction, creates the λ_p-period oscillations as a function of the tip–edge separation L' in the images.

To model the scattering-type SNOM images shown in Extended Data Fig. 2a quantitatively, we used an electromagnetic solver described previously[9,32]. This software uses a boundary-element method to solve the integral equation of the convolution type

$$\hat{\epsilon}\Phi = \Phi_{tip} \tag{1}$$

for the quasi-static scalar potential $\Phi(x, y)$ induced in graphene in response to the external perturbation, which comprises the incident field plus the potential $\Phi_{tip}(x, y)$ created by the charges on the tip (the hat denotes an integral operator). The potential Φ_{tip} is not known a priori; it depends on Φ, on the incident field and the tip geometry in a self-consistent manner (see refs [9,32]). In our model, the tip is approximated by a metallic spheroid[9,15] and the incident field is taken to be uniform.

The kernel of equation (1) has the physical meaning of the two-dimensional dielectric function of graphene, which has the well known momentum-space representation

$$\epsilon(q, \omega) = 1 - \frac{|q|}{q_p(\omega)}$$

where

$$q_p(\omega) = q_p'(\omega) + iq_p''(\omega) = \frac{i\kappa\omega}{2\pi\sigma(\omega)} \quad \text{and} \quad \sigma(\omega) = \sigma'(\omega) + i\sigma''(\omega) \tag{2}$$

are the complex plasmon wavenumber and the sheet conductivity of graphene, respectively, and κ is the effective dielectric function of the graphene environment; see 'Plasmonic damping rate analysis'. We note that the plasmon wavelength λ_p and the plasmon quality factor Q are given by

$$\lambda_p = \frac{2\pi}{q_p'} \quad \text{and} \quad Q = \frac{q_p'}{q_p''} \tag{3}$$

Our software code solves equation (1) and the equations relating Φ_{tip} and Φ numerically and subsequently computes the observable scattering-type SNOM signal as a function of the tip–edge distance.

To fit the experimental data with these simulations, we adopted the Drude model of the conductivity

$$\sigma(\omega) = \frac{i}{\pi}\frac{D}{\omega + i\gamma(\omega)} \tag{4}$$

where the Drude weight

$$D = \frac{2e^2}{\hbar^2} T \ln\left[2\cosh\left(\frac{\mu}{2T}\right)\right] \simeq \frac{e^2}{\hbar^2}|\mu| \tag{5}$$

and the scattering rate γ are adjustable parameters. In equation (5), $\hbar$ is the Planck constant and μ is the chemical potential. Our best fits of the computed line profiles to the experimental data are shown in Extended Data Fig. 2b. The physical origin of γ is discussed in Methods section 'Plasmonic damping rate analysis'.

Gate dependence of the plasmon fringes. Extended Data Fig. 3 illustrates the gating-induced carrier-density dependence of nanoscale infrared images at cryogenic temperature. These data are presented in a form of two-dimensional map of the near-field amplitude s as a function of the gate voltage and the distance L from the Au launcher. These data were acquired by repeated scanning along the same linear path for a set of increasing gate voltages. The direction of the scan was normal to the Au launcher. The data shown in Extended Data Fig. 3 encode information about the variation of λ_p and γ with electrostatic doping. For example, the period of the plasmon fringes, which is equal to λ_p, is seen to monotonically increase with V_g. The procedure used for the quantitative determination of λ_p and γ is detailed in 'Simulations of plasmons launched by the Au pad'.

The optical constants of hBN. The optical constants of hBN in infrared frequencies are governed by its optical phonon modes. To the best of our knowledge, there are no previously published data on the hBN phonon parameters at low temperatures. We have obtained this information by measuring the reflectance $R(\omega)$ of exfoliated hBN micro-crystals on SiO_2/Si substrates. Because of the limited sample area, these experiments were performed using a diffraction-limited infrared microscope, which imposed a low-frequency cut-off for these far-field reflectance measurements. These experiments were done under high vacuum conditions. Extended Data Fig. 4 shows the optical image of the hBN flake under study.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

The flake thickness is ~17 nm, which is close to the thickness of the bottom hBN layer in our hBN/graphene/hBN device structures. Extended Data Fig. 5 shows the $R(\omega)$ value of hBN/SiO_2/Si multilayer structures at different temperatures normalized to that of a gold reference. As one can see, $R(\omega)$ is dominated by two resonance peaks, one at ~1,070 cm^{-1} due to SiO_2 phonons and the other at ~1,370 cm^{-1} due to hBN in-plane phonons. Extraction of the phonon parameters was done with the help of the RefFIT package[33], which fits $R(\omega)$ spectra using the multi-layer Fresnel reflection formulas, with every layer modelled with Lorentzian oscillators. The response of amorphous SiO_2 on a Si substrate is complex and, according to earlier studies, can be described with multiple Lorentzians as

$$\epsilon_2(\omega) = \epsilon_\infty\left(1 + \sum_j \frac{\omega_{pj}^2}{\omega_{tj}^2 - \omega^2 - i\gamma_j\omega}\right) \quad (6)$$

where ω_{tj} is the transverse optical phonon frequency of the phonon species and ϵ_∞ is the high-frequency dielectric constant. The parameters ω_{tj}, ω_{pj} and γ_j for the room-temperature case are given in Extended Data Table 1. The corresponding fits to the reflectance data are shown by the red dashed lines in Extended Data Fig. 5. For the in- and out-of-plane permittivities of hBN we found it sufficient to use the single-Lorentzian model

$$\epsilon_x(\omega) = \epsilon_{\infty x}\left(1 + \frac{\omega_{px}^2}{\omega_{tx}^2 - \omega^2 - i\gamma_x\omega}\right)$$

$$\epsilon_z(\omega) = \epsilon_{\infty z}\left(1 + \frac{\omega_{pz}^2}{\omega_{tz}^2 - \omega^2 - i\gamma_z\omega}\right)$$

Extended Data Fig. 6 shows the obtained linewidth γ_x and frequency ω_{tx} of the in-plane phonon. In particular, the linewidth is found to vary linearly with T, decreasing from ~9 cm^{-1} at room temperature to ~5 cm^{-1} at cryogenic temperatures. This temperature dependence has been included in our theoretical estimation of the dielectric losses (Methods section 'Plasmonic damping rate analysis'). Unfortunately, sample thickness prohibits the evaluation of the out-of-plane permittivity of hBN from these $R(\omega)$ results. Instead, we employed the Fourier-transform infrared nanospectroscopy (nano-FTIR) technique to determine ϵ_z (see Methods section 'Plasmonic damping rate analysis' and Extended Data Table 2).

Simulations of plasmons launched by the Au pad. The electromagnetic solver described in 'Gate dependence of the plasmon fringes' gives an accurate representation of the total near-field signal near the edge of graphene. However, it is more difficult to implement the same approach for a system that contains three conductors, each of a nontrivial shape: a graphene micro-crystal, an elongated tip and a box-like Au pad. Fortunately, this is not necessary. For the purpose of extracting the plasmon wavelength λ_p and the quality factor Q, we only need to consider the long-range plasmonic field launched by the pad. We assume that in the interior of the graphene flake, the tip acts as a passive detector and that the near-field signal that it registers is directly proportional to the normal component E^z of the electric field underneath the tip. The first assumption is reasonable if the pad is a more efficient plasmon launcher than the tip. This is corroborated by the fact that the plasmonic fringes are dominated by simple λ_p-period oscillations instead of 'double peaks', as in Extended Data Fig. 2. The second assumption is not strictly true, but it is very common in the literature[8,9,32]. On the basis of these considerations, we included only the Au pad and graphene in our model.

The simulation was performed using a publicly available software tool[34] that implements the boundary-element method to solve electrodynamic problems involving nanoparticles. In our simulations the pad's dimensions were 1,000 nm × 500 nm × 60 nm. The last two values (the width and thickness) matched the physical ones. However, the first dimension (the length of the pad in the x direction) was shorter than that of the actual device because we found it computationally prohibitive to simulate such a many-micrometre-long object on our computer (a desktop with a memory of 16 GB and a clock speed of 2.7 GHz). We believe that this reduction in the length of the pad in our model is still reasonable because the pad's length controls mostly the antenna effect (that is, the overall magnitude of the field at its leading edge), whereas we want to reproduce the profile of the field distribution—not its absolute magnitude. We also rounded the pad's corners using a curvature radius of 60 nm to obtain a smaller boundary-element method mesh (see Extended Data Fig. 7a) and speed up the computation.

The permittivity of Au in our model was taken to be $\epsilon_{Au} = (1.0 + 1.0i) \times 10^3$ but the results were insensitive to this parameter as long as it was large enough by absolute value. We modelled graphene as an infinite planar slab of thickness $d_g = 0.1$ nm and three-dimensional permittivity $\epsilon_g = -\coth(q_p d_g/2)$, which is related to the two-dimensional dielectric function introduced in Methods section 'Plasmons near a graphene edge' as $\epsilon_g \approx 2\epsilon(q,\omega)/(qd_g)$. This relation ensures that the slab supports propagating waves of momentum q_p. The results for the quantity of interest, which is the electric field just above the graphene, were practically independent of the arbitrary parameter d_g as long as d_g was small enough. The hBN substrate was not explicitly included in the simulation, but its effect was accounted for implicitly by using the proper value of q_p. The external field was taken to be uniform and linearly polarized in the y direction, Extended Data Fig. 7a. Using this model, we numerically calculated the field component E^z within a 4,000-nm^2 square region just above the graphene layer. An example of such a two-dimensional map is shown in the false-colour image of Extended Data Fig. 7b for the simulation parameters $\lambda_p = 170$ nm and $Q = 130$, which are the ones that provide the best fit to the experimental data obtained at $V_g = 75$ V and $T = 60$ K. The qualitative features are similar to those seen in Fig. 2. Namely, the plasmon oscillations display a generic wave pattern that starts as a plane wave near the Au launcher, then turns into a Fresnel diffraction pattern further away and afterwards gradually transforms into a decaying cylindrical wave modulated by the Fraunhofer diffraction features; see Extended Data Fig. 7b. To avoid complications in the fitting procedure by these plasmonic diffraction effects and to suppress the influence of statistical noise in the experimental data, we averaged multiple line traces from both the simulation and the experimental two-dimensional maps over the strip $x = -300$ nm to $x = 300$ nm. This averaging gave us smooth line traces suitable for fitting. Our best fit for the above values of V_g and T is illustrated in Extended Data Fig. 7c. The red trace is the data and the blue trace is the calculated, strip-averaged E^z. The same fit can be also found in Fig. 2, together with those for the remaining data. To align the phases of the two curves in Extended Data Fig. 7c, we used a phase shift θ. However, this additional adjustable parameter does not affect our main goal of extracting Q, which we achieved by analysing the decay of the absolute value of E^z, that is, the envelope of the oscillations. The fit in Extended Data Fig. 7c is clearly good, as shown in Fig. 2. Hence, in our subsequent analysis we relied on the deduced Q, as described in 'Electron scattering by the acoustic phonons of graphene'.

Plasmonic damping rate analysis. *Losses in the dielectric environment of grapheme.* According to equations (2) and (3), the reciprocal of the plasmonic quality factor Q can be split into two additive components

$$Q^{-1} = \frac{q_p''}{q_p'} \approx \frac{\sigma'}{\sigma''} + \frac{\kappa''}{\kappa'}$$

In this section we discuss the second term, which represents losses in the dielectric environment of graphene. To evaluate such losses that are inherent to our encapsulated structures, we need to find the effective permittivity $\kappa = \kappa'(q,\omega) + \kappa''(q,\omega)$. This requires solving Maxwell's equations for the potential distribution in our multilayer structure (Extended Data Fig. 8a), which is created in response to a periodically modulated charge density of $\sim e^{i\boldsymbol{q}\cdot\boldsymbol{r}}$ introduced in the graphene plane, where $\boldsymbol{r}$ is the two-dimensional coordinate on the graphene plane. In the quasi-static limit of large momenta, $q \gg \omega/c$, the solution of this electrodynamic problem is straightforward, so we present only the final result

$$\kappa(q,\omega) = \sqrt{\epsilon_x}\sqrt{\epsilon_z}\frac{1 - l_1 l_2}{(1 + l_1)(1 + l_2)} \quad (7)$$

where

$$l_j = r_j e^{2ik_z d_j}, \quad k_z = iq\frac{\sqrt{\epsilon_x}}{\sqrt{\epsilon_z}} \quad \text{and} \quad r_j = \frac{\epsilon_z k_z - i\epsilon_j q}{\epsilon_z k_z + i\epsilon_j q} \quad (8)$$

It is instructive to consider two limiting cases. If $d_1, d_2 \to \infty$, then $e^{2ik_z d_j} \to 0$, so the effective permittivity $\kappa_{hBN} = \sqrt{\epsilon_x}\sqrt{\epsilon_z}$ of bulk hBN is obtained, as expected. In the opposite limit, $d_1, d_2 \to 0$, we find $\kappa = (\epsilon_1 + \epsilon_2)/2$, that is, the effective permittivity is equal to the average of the permittivities of the semi-infinite top and bottom layers (vacuum and SiO_2), whereas the permittivity of hBN does not enter the calculation. Our device had $d_1 = 7$ nm and $d_2 = 14$ nm (Extended Data Fig. 8a). In this intermediate case, it is imperative to properly account for contributions from both hBN and SiO_2 losses.

As mentioned in 'The optical constants of hBN', the temperature dependence of the damping constant γ_z of the hBN c-axis permittivity is not experimentally attainable (where the c axis is the high-symmetry axis perpendicular to the a–b plane). Owing to the thermal occupation factors involved, both γ_x and γ_z are expected to monotonically decrease with temperature. We thus assume that γ_z has the same temperature dependence as γ_x, that is,

$$\gamma_z(T) = \gamma_x(T)\frac{\gamma_z(300\,\mathrm{K})}{\gamma_x(300\,\mathrm{K})} \quad (9)$$

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

To determine $\gamma_z(300\,\mathrm{K})$, we performed nano-FTIR experiments for an hBN crystal taken away from the sample edges (Extended Data Fig. 8b) from which we obtained $\gamma_z(300\,\mathrm{K}) \approx 3.4\,\mathrm{cm}^{-1}$ can be obtained via numerical modelling, in agreement with previous work[35]. For the plasmon wavelength $\lambda_p = 170\,\mathrm{nm}$, the final estimate of the environmental dielectric loss

$$\gamma_{\mathrm{env}}(\omega) \equiv \omega \frac{\kappa''(\omega)}{\kappa'(\omega)}$$

calculated with the help of equations (7)–(9), is plotted in Extended Data Fig. 8c for $\gamma_z(300\,\mathrm{K}) = 3.4\,\mathrm{cm}^{-1}$. Clearly, $\gamma_{\mathrm{env}}(T)$ is well approximated by a linear function.

Electron scattering by the acoustic phonons of grapheme. The scattering rate γ was defined in equation (4) as one of the two parameters (along with the Drude weight, D, in equation (5)) that define the graphene conductivity $\sigma = \sigma' + \sigma''$. In this section, we discuss theoretical estimates of this scattering rate, which is generally temperature-dependent. In this discussion, the conductivity appears as a primary quantity and γ as a derived one. In the case of interest, $\gamma \ll \omega$, it is given by

$$\gamma(\omega) = \omega \frac{\sigma'}{\sigma''} = \frac{\pi}{D}\omega^2 \sigma'(\omega) \tag{10}$$

In principle, the conductivity and the scattering rate depend on both the frequency, ω, and the momentum, q. The characteristic scale of the momentum dependence is ω/v_F, where v_F is the Fermi velocity. The plasmon momentum, $q_p \ll \omega/v_F$, is much smaller. Therefore, to calculate the plasmonic losses it is sufficient to compute the zero-momentum scattering rate $\gamma(\omega) \equiv \gamma(q=0,\omega)$, which is an easier task. Noting that the power P absorbed per unit area of the system in the presence of an in-plane electric field $Ee^{-i\omega t} + E^* e^{i\omega t}$ is equal to $P = 2\sigma'(\omega)|E|^2$, the desired scattering rate can be expressed as

$$\gamma(\omega) = \frac{\pi\omega^2}{2D}\frac{P}{|E|^2} = \frac{\omega^2}{e^2 v_F^2}\frac{\partial\mu}{\partial n}\frac{P}{|E|^2} \tag{11}$$

We begin with the acoustic phonon contribution to γ, which we denote γ_A. The velocities of the longitudinal and transverse acoustic phonons in graphene, $v_l = 2.2\times10^6\,\mathrm{cm\,s^{-1}}$ and $v_t = 1.4\times10^6\,\mathrm{cm\,s^{-1}}$, are much smaller than the Fermi velocity, $v_F \approx 1.0\times10^8\,\mathrm{cm\,s^{-1}}$. This implies that the perturbations created by acoustic phonons are quasi-static for electrons. In conventional metals and semiconductors, such perturbations are described in terms of the deformation potential. However, in graphene, it has become common to refer only to a part of the total perturbation as the deformation potential, namely, the part that is proportional to the diagonal component of the strain. The remaining part is referred to as the gauge field or the pseudo-magnetic field[36]. The physical picture that goes with this terminology is that the deformation potential shifts the Dirac cones of graphene in energy, whereas the pseudo-magnetic field moves them in the momentum space, in opposite directions for the two valleys. The issue of the relative importance of deformation potential and pseudo-magnetic field remains unsettled. Although the deformation potential is absent[37] or very small[38] in a naive tight-binding model, it is nevertheless often assumed to be the dominant contribution[18,19,39,40]. However, recent first-principles calculations[20,41] found that the deformation potential is indeed negligible. We assume this to be the case and consider pseudo-magnetic field scattering only. This permits us to neglect electron–electron interactions because the pseudo-magnetic field, unlike the deformation potential, does not perturb the electron density.

To find the phonon-induced correction to the power absorption P, we use the Fermi golden rule (see Extended Data Fig. 9)

$$\begin{aligned} P \;=\; & 2\pi\omega\left|\frac{eA}{c}\right|^2 \int \frac{N\mathrm{d}^2k}{(2\pi)^2}\frac{\mathrm{d}^2q}{(2\pi)^2} \\ & \times \sum_{\zeta,\zeta_\gamma}\sum_{S,S'}\sum_{\nu=l,t} \zeta\zeta_\gamma f(\varepsilon_{S\boldsymbol{k}})[1-f(\varepsilon_{S'\boldsymbol{k}'})]\left[n_B\left(\zeta\hbar\omega_{\nu\boldsymbol{q}}\right)+1\right] \\ & \times \delta\left(\varepsilon_{S'\boldsymbol{k}'} - \varepsilon_{S\boldsymbol{k}} + \zeta\hbar\omega_{\nu\boldsymbol{q}} - \zeta_\gamma\hbar\omega\right) \\ & \times \left|\sum_{S''}\left(\frac{W^{\nu}_{S'\boldsymbol{k}';S''\boldsymbol{k}}v^x_{S''\boldsymbol{k};S\boldsymbol{k}}}{\varepsilon_{S\boldsymbol{k}} - \varepsilon_{S''\boldsymbol{k}} + \zeta_\gamma\hbar\omega} - \frac{v^x_{S'\boldsymbol{k}';S''\boldsymbol{k}'}W^{\nu}_{S''\boldsymbol{k}';S\boldsymbol{k}}}{\varepsilon_{S'\boldsymbol{k}'} - \varepsilon_{S''\boldsymbol{k}'} + \zeta_\gamma\hbar\omega}\right)\right|^2 \end{aligned} \tag{12}$$

Here $A = cE/(i\omega)$ is the in-plane component of the vector potential; c is the speed of light, $N = 4$ is the spin–valley degeneracy; $\zeta = 1$ ($\zeta_\gamma = -1$) represents the emission of a phonon (photon) and $\zeta = -1$ ($\zeta_\gamma = 1$) represents the absorption of a phonon (photon); $\varepsilon_{S\boldsymbol{k}} = S\hbar v_F|\boldsymbol{k}|$ is the energy of an electron with momentum $\boldsymbol{k}$ and band index $S = \pm1$; $\omega_{\nu\boldsymbol{q}} = v_\nu|\boldsymbol{q}|$ is the frequency of the phonon of type ν and momentum $\boldsymbol{q}$; $f(\varepsilon)$ and $n_B(\varepsilon)$ are the Fermi and Bose distribution functions, respectively; and $\boldsymbol{k}' = \boldsymbol{k} - \boldsymbol{q}$. The matrix elements $W^{\nu}_{S\boldsymbol{k};S'\boldsymbol{k}'} = \psi^\dagger_{S\boldsymbol{k}} W^{\nu}\psi_{S'\boldsymbol{k}'}$ of the electron–phonon coupling and the electron velocity $v^x_{S\boldsymbol{k};S'\boldsymbol{k}} = \psi^\dagger_{S\boldsymbol{k}} v^x \psi_{S'\boldsymbol{k}}$ are calculated from[41]

$$\begin{aligned} W^{\nu} &= i\beta_A\sqrt{\frac{\hbar q^2}{\mu_s\omega_{\nu\boldsymbol{q}}}}\begin{pmatrix} 0 & i^{s_\nu}e^{2i\theta_{\boldsymbol{q}}} \\ i^{-s_\nu}e^{-2i\theta_{\boldsymbol{q}}} & 0\end{pmatrix} \\ \nu^x &= \nu_F\begin{pmatrix}0&1\\1&0\end{pmatrix} \\ \psi_{S\boldsymbol{k}} &= \frac{1}{\sqrt{2}}\begin{pmatrix} e^{-i\theta_{\boldsymbol{k}}} \\ S\end{pmatrix} \end{aligned} \tag{13}$$

where $\beta_A = 5.0\,\mathrm{eV}$ is the pseudo-magnetic field coupling constant, $\mu_s = 7.6\times10^{-8}\,\mathrm{g\,cm^{-2}}$ is the graphene mass density, s_ν is equal to 0 (1) for $\nu = l$ ($\nu = t$) phonons and $\theta_{\boldsymbol{q}}$ is the polar angle of $\boldsymbol{q}$, with the x axis being along a zigzag direction.

Numerical evaluation of $\gamma_A(\omega, T)$ using equations (11)–(13) for $\omega = 886\,\mathrm{cm^{-1}}$ and for the d.c. limit $\omega \to 0$ gives the results shown in Extended Data Fig. 10a by the green and blue lines, respectively. These lines are almost straight and nearly coincide with one another in the range of our experiments, $T \geq 60\,\mathrm{K}$. For the d.c. scattering rate $\gamma_A(\omega = 0, T)$, such a linear behaviour with respect to T is expected[18–20,37–39,41]. γ_A can be derived analytically by assuming $|\mu| \gg \omega, T$ and $T \gg T_{BG}$, where $T_{BG} = 2|\mu|v_l/v_F$ is the Bloch–Grüneisen temperature. This set of inequalities permits one to neglect $\omega_{\nu\boldsymbol{q}}$ inside the δ-function and to replace $f(\varepsilon)$ by a step-function and $n_B(\varepsilon)$ by T/ε. After some algebra, we find

$$\begin{aligned} \gamma_A(0) &= \frac{1}{\hbar^3}\frac{\beta_A^2|\mu|}{\mu_s v_F^2}\left(\frac{1}{v_l^2}+\frac{1}{v_t^2}\right)T \\ \gamma_A(\omega) &\approx \left(1+\frac{1}{12}\frac{\omega^2}{\mu^2}\right)\gamma(0) \end{aligned} \tag{14}$$

The first equation in equation (14) is in agreement with prior work[20,38–41], as is also the common observation that the linear temperature dependence sets in already at $T \approx 0.2T_{BG} \approx (10\,\mathrm{K})\sqrt{n[10^{12}\,\mathrm{cm^{-2}}]} \approx 25\,\mathrm{K}$; see Extended Data Fig. 10. The second equation in equation (14) is our new result. Importantly, it explains why the infrared scattering rate $\gamma_A(\omega, T)$ at the frequency of interest, $\omega = 886\ \mathrm{cm^{-1}} \approx 0.4|\mu|$, is hardly different from the d.c. rate $\gamma_A(0, T)$ in Extended Data Fig. 10a.

The weak frequency dependence given by our formula is inconsistent with the rapid increase of $\gamma_A(\omega, T)$ with ω obtained in a previous theoretical work[18] that studied electron–phonon scattering by the deformation potential mechanism. By examining the formulas given in ref. [18], we traced the difference between the deformation potential and pseudo-magnetic field cases to a strongly ω-dependent electron screening. This screening influences the deformation potential in electron−phonon coupling but plays no role in the pseudo-magnetic field channel we consider here.

At temperatures $T \lesssim 0.2T_{BG}$ the assumptions used in deriving equation (14) do not hold. Such temperatures have not been reached in our experiments, but we can discuss the predictions of our theory. In this ultimate low-temperature regime, the d.c. and alternating current (a.c.) scattering rates would no longer have the same scaling with temperature. The d.c. rate (Extended Data Fig. 10a, blue line) should vanish, whereas the a.c. one (Extended Data Fig. 10a, green line) should approach a non-zero value as $T \to 0$. The physical reason for this difference is that the scattering rate $\gamma_A(\omega, 0)$ at absolute zero is solely due to emission of acoustic phonons, which is possible only at a finite ω.

The red and black solid lines in Extended Data Fig. 10a represent the electron–phonon inter-valley scattering rate. This process, as well as the total electron–phonon scattering rates plotted in Extended Data Fig. 10b and c, are discussed in the next section.

Inter-valley electron-phonon scattering. Inter-valley electron–phonon scattering is dominated by transverse optical phonons at the K-point of the Brillouin zone, the so-called A_1' phonons[42]. The matrix element involved in this scattering

$$W^{K}_{S\boldsymbol{k}S'\boldsymbol{k}'} = i\beta_K\sqrt{\frac{2\hbar}{\mu_s\omega_K}}\begin{pmatrix} e^{i\theta_k} \\ s\end{pmatrix}^\dagger\begin{pmatrix}0 & i\\ -i & 0\end{pmatrix}\begin{pmatrix} e^{-i\theta_{k'}} \\ s'\end{pmatrix} \tag{15}$$

is given by[41]

where β_K is the electron–A_1'-phonon coupling constant. According to previous calculations[20,42] of the corresponding d.c. scattering rate $\gamma_K(0)$, A_1'-phonon scattering is activated at $T \approx 0.15\hbar\omega_K \approx 250$ K, and at room temperature it is responsible for a large fraction of the total $\gamma(0) \approx \gamma_A(0) + \gamma_K(0)$. Here we show that inter-valley scattering has an even larger contribution $\gamma_K(\omega)$ to the infrared scattering rate and that it begins at a temperature as low as 100 K. This scattering mechanism can therefore naturally explain the observed super-linear temperature increase of the plasmon damping seen in Extended Data Fig. 10.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Within a naive tight-binding model[20,37], one finds $\beta_K = 4\beta_A/b = 14\,\text{eV Å}^{-1}$, where $b = 1.42$ Å is the carbon–carbon bond length. However, virtual electron–phonon scattering, such as electron screening and vertex corrections, are important for the inter-valley processes. These many-body effects cause a red shift of the A_1'-phonon frequency ω_K (Kohn anomaly[42]) and an enhancement[41–43] of the electron–phonon coupling constant β_K. Both effects are difficult to compute accurately, and so we treat ω_K and β_K as adjustable parameters.

To calculate γ_K we use again the Fermi golden rule, equation (12), this time with the matrix element from equation (15). For high doping, $|\mu| \gg \omega$, we can keep only the intraband terms $S'' = S' = S$, which are the dominant ones. After some algebra, we arrive at

$$\gamma_K(\omega) = \frac{3}{2}\frac{\beta_K^2|\mu|}{\mu_s\omega_K\omega\hbar^3 v_F^2}\sum_{\zeta,\zeta_\gamma}\zeta\zeta_\gamma\int_0^\infty x(|\mu|x - \zeta\hbar\omega_K + \zeta_\gamma\hbar\omega) \times f(\mu x)[1 - f(|\mu|x - \zeta\hbar\omega_K + \zeta_\gamma\hbar\omega)][1 + n_B(\zeta\hbar\omega_K)]\mathrm{d}x$$

This integral can be readily evaluated numerically[34]. At low temperatures the growth of $\gamma_K(\omega) \approx \exp[(\omega_K - \omega)/T]$ is roughly exponential because it is dictated by the thermal tails of the occupation factors of the phonons and electrons. This consideration makes it clear why the infrared scattering rate has an earlier onset in T than the d.c. one. In Extended Data Fig. 10c we show the results of our calculations for the total rate $\gamma = \gamma_A + \gamma_K$. The circles (squares) represent the plasmon damping rates extracted from the experiment, with the environmental loss rate chosen as in Extended Data Fig. 8c. For both d.c. and infrared cases, shown by the red and blue curves, respectively, good agreement between the theory and the experimental data is achieved for $\beta_K = 14\,\text{eV Å}^{-1}$ and $\omega_K = 1,200\,\text{cm}^{-1}$. Note that these values are basically the same as the previous first-principles calculations results[20,41,42], where $\beta_K = 13.9\,\text{eV Å}^{-1}$ and $\omega_K = 1,209\,\text{cm}^{-1}$. Other parameters that we used were $\mu = 0.27\,\text{eV}$ and $\beta_A = 5.0\,\text{eV}$.

Electron–electron scattering. In this section we estimate the plasmon damping due to electron–electron interactions using an original method. For previous work on this subject see, for example, ref. [43]. In general, electron interactions can influence the conductivity only if the Galilean invariance of the system is broken[44], which is a weak effect in doped graphene at low temperatures.

We begin by noting that the electron–electron collision rate[44]

$$\Gamma_{ee} = \frac{\pi}{4}\frac{T^2}{|\mu|}\ln\left(\frac{2|\mu|}{T}\right) \tag{16}$$

is comparable to the phonon-induced scattering rate; see Extended Data Figs. 10 and 11. However, such collisions conserve the total momentum and thus approximately conserve the total current as well. In fact, there exists a nonequilibrium, current-carrying state that is immune to electron–electron collisions, namely, a Fermi sea with a Doppler-shifted energy spectrum $\varepsilon_{S\boldsymbol{k}} \to \varepsilon_{S\boldsymbol{k}} - \hbar u^x k^x$. This state corresponds to a uniform hydrodynamic flow with velocity u^x. In comparison, the Fermi distribution of noninteracting electrons subject to an electric field E is typically shifted in momentum as $\hbar k^x \to \hbar k^x - \Delta p^x$ with $\Delta p^x = eE\gamma^{-1}$. At low temperatures, this latter state is similar, but not identical, to a hydrodynamic flow with velocity $u^x = \Delta p^x/m$, where $m = \hbar k_F/v_F$ is the effective mass and k_F is the Fermi momentum. We can call these two types of perturbed distribution functions the velocity mode and the momentum mode, respectively. The Galilean invariance is indeed broken because m depends on the electron energy. However, this dependence is weak; therefore, we can anticipate that the damping rate entering the conductivity is much smaller than the electron collision rate Γ_{ee}.

To proceed, we suppose that both momentum-conserving and nonconserving processes are present, with rates Γ_{ee} and Γ_d, respectively. As shown in recent work[22], the form of the conductivity can be obtained by assuming that the perturbation of the electron distribution function is a linear combination of the velocity and momentum modes. This ansatz leads to the two-term Drude formula

$$\sigma(\omega) = \frac{i}{\pi}\frac{D_h}{\omega + i\Gamma_d} + \frac{i}{\pi}\frac{D - D_h}{\omega + i\Gamma_d + i\Gamma_{ee}}$$

(see also ref. [45]) in lieu of the conventional single-term one, equation (4). We can still define the effective electron–electron scattering rate similarly to equation (10) to obtain

$$\gamma_{ee} \equiv \frac{\pi}{D}\omega^2\sigma'(\omega) - \Gamma_d \approx \frac{D - D_h}{D}\Gamma_{ee}, \quad \omega \gg \Gamma_{ee}, \Gamma_d \tag{17}$$

Here $D = \int_{-\infty}^{\infty}\sigma'(\omega)\mathrm{d}\omega$ is the total optical weight given by equation (4) and $D_h = \pi e^2 n/m_h$ is the hydrodynamic Drude weight. For weak interactions, the corresponding hydrodynamic mass m_h is given by[22,46]

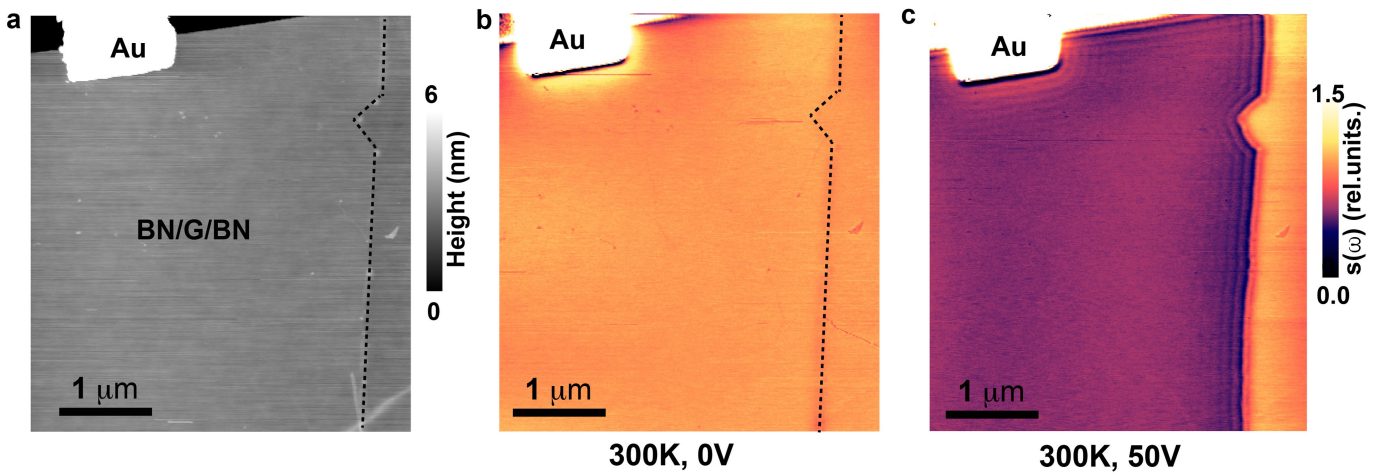
$$\begin{aligned} m_h &= \frac{3}{2}\frac{1}{nv_F^2}\sum_{S=\pm1}\left\{\int_{-\infty}^{\infty}\frac{N\mathrm{d}^2k}{(2\pi)^2}[f(\varepsilon_{S\boldsymbol{k}}) - \Theta(-S)]\varepsilon_{S\boldsymbol{k}}\right\} \\ &= \frac{3}{\pi}\frac{T^3}{\hbar^2 v_F^4 n}\left[\frac{\pi^2}{3}\frac{\mu}{T} + \frac{1}{3}\frac{\mu^3}{T^3} - 4\,\mathrm{Li}_3(-e^{\mu/T})\right] \end{aligned} \tag{18}$$

where Θ is the unit step function and $\mathrm{Li}_n(z)$ is the polylogarithm function of order n and argument z. The crucial point is that at the temperatures of interest, $T \ll |\mu|$, the two Drude weights, D and D_h, are nearly equal. As a result, the plasmon damping rate γ_{ee} due to electron interactions is indeed much smaller than the electron collision rate Γ_{ee}; see equation (17) and Extended Data Fig. 11.

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

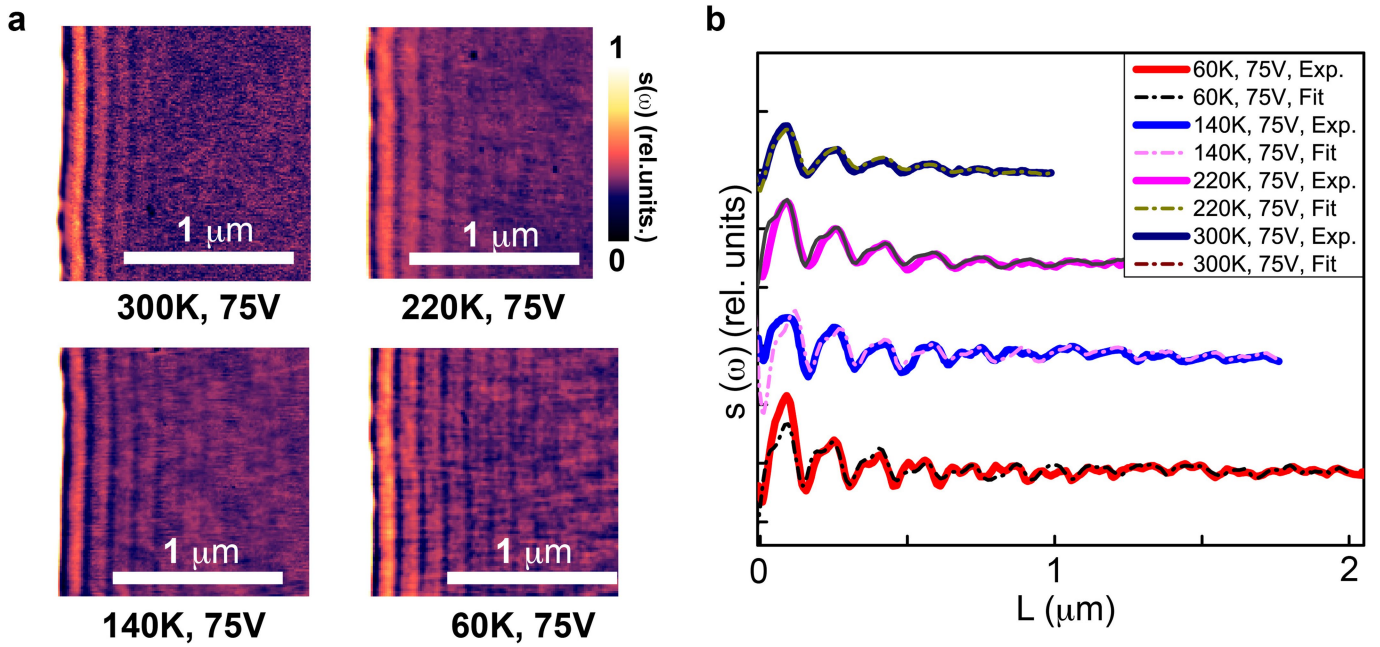
32. Fei, Z. et al. Infrared nanoscopy of Dirac plasmons at the graphene–SiO_2 interface. *Nano Lett.* **11**, 4701–4705 (2011).
33. Kuzmenko, A. B. RefFIT: software to fit optical spectra, v1.2.95.
34. Waxenegger, J., Trügler, A. & Hohenester, U. Plasmonics simulations with the MNPBEM toolbox: consideration of substrates and layer structures. *Comput. Phys. Commun.* **193**, 138–150 (2015).
35. Dai, S. et al. Tunable phonon polaritons in atomically thin van der Waals crystals of boron nitride. *Science* **343**, 1125–1129 (2014).
36. Amorim, B. et al. Novel effects of strains in graphene and other two-dimensional materials. *Phys. Rep.* **617**, 1–54 (2016).
37. Pietronero, L., Strässler, S., Zeller, H. R. & Rice, M. J. Electrical conductivity of a graphite layer. *Phys. Rev. B* **22**, 904–910 (1980).
38. Woods, L. M. & Mahan, G. D. Electron-phonon effects in graphene and armchair (10,10) single-wall carbon nanotubes. *Phys. Rev. B* **61**, 10651 (2000).
39. von Oppen, F., Guinea, F. & Mariani, E. Synthetic electric fields and phonon damping in carbon nanotubes and graphene. *Phys. Rev. B* **80**, 075420 (2009).
40. Mariani, E. & von Oppen, F. Temperature-dependent resistivity of suspended graphene. *Phys. Rev. B* **82**, 195403 (2010).
41. Sohier, T. et al. Phonon-limited resistivity of graphene by first-principles calculations: electron–phonon interactions, strain-induced gauge field, and Boltzmann equation. *Phys. Rev. B* **90**, 125414 (2014).
42. Piscanec, S., Lazzeri, M., Mauri, F., Ferrari, A. C. & Robertson, J. Kohn anomalies and electron–phonon interactions in graphite. *Phys. Rev. Lett.* **93**, 185503 (2004).
43. Basko, D. M. & Aleiner, I. L. Interplay of Coulomb and electron–phonon interactions in graphene. *Phys. Rev. B* **77**, 041409 (2008).
44. Giuliani, G. & Vignale, G. *Quantum Theory of the Electron Liquid* (Cambridge Univ. Press, Cambridge, 2005).
45. Maslov, D. L. & Chubukov, A. V. Optical response of correleted electron systems. *Rep. Prog. Phys.* **80**, 026503 (2017).
46. Sun, Z., Basov, D. N. & Fogler, M. M. Adiabatic amplification of plasmons and demons in 2D systems. *Phys. Rev. Lett.* **117**, 076805 (2016).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



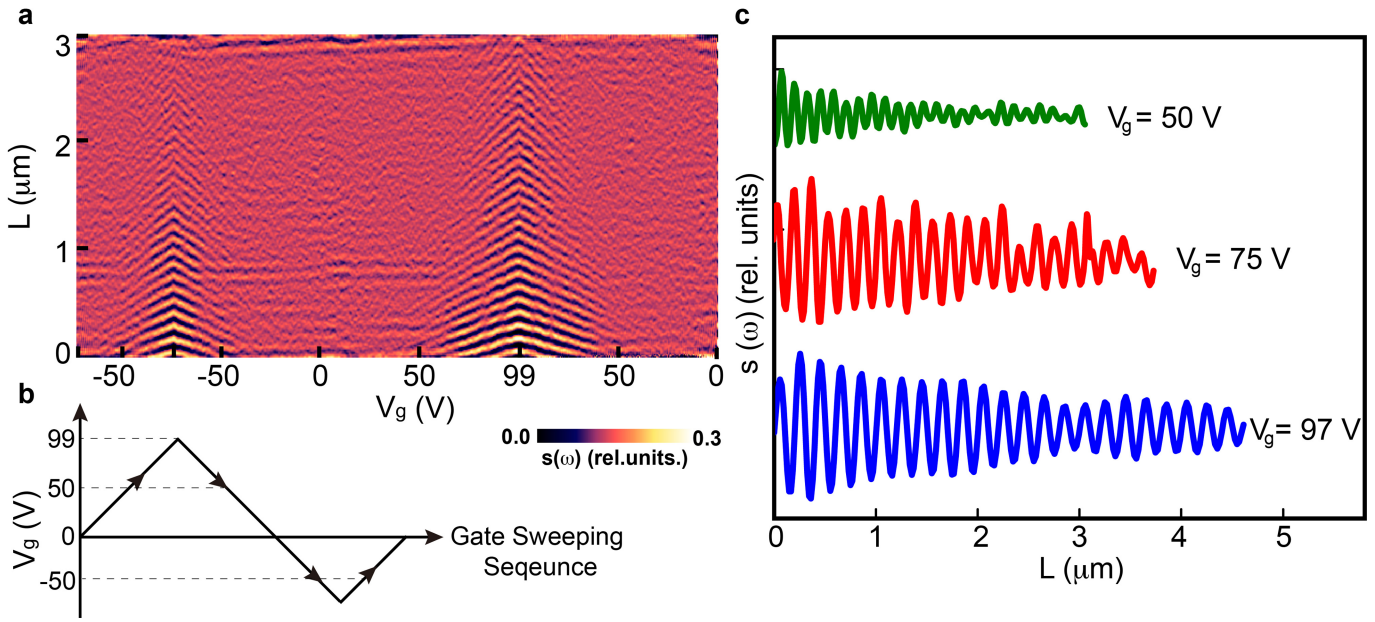
Extended Data Fig. 1 | AFM topography and $s(\omega)$ image of the hBN/graphene/hBN encapsulated device. a, AFM topography image. The black dashed line marks the physical edge of graphene. **b**, **c**, Room-temperature $s(\omega)$ images obtained at $V_g = 0$ V and $V_g = 50$ V, respectively. G, graphene.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



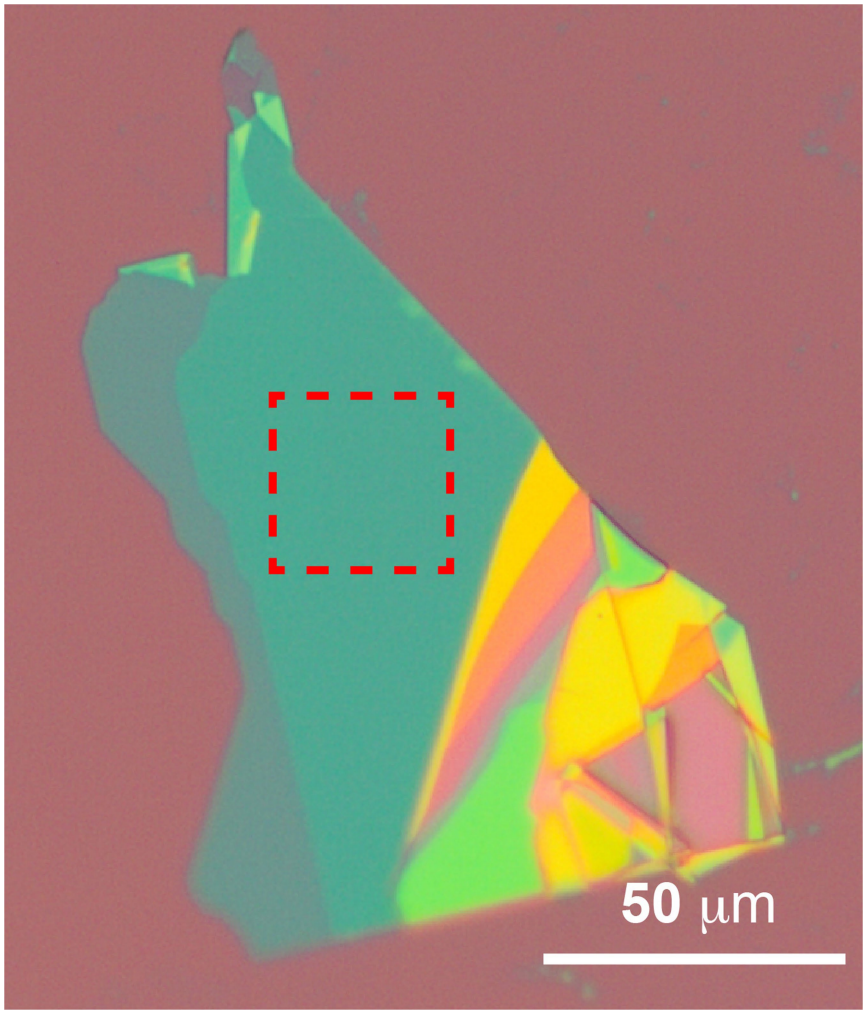
Extended Data Fig. 2 | Plasmonic fringes near a graphene edge at different temperatures. a, Maps of the near-field amplitude s measured at several temperatures for fixed $V_g = 75$ V and $\omega = 886\,\mathrm{cm}^{-1}$. **b**, Line profiles obtained by averaging over the vertical coordinates in each map. The solid lines are the data and the dash-dotted lines are the simulations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



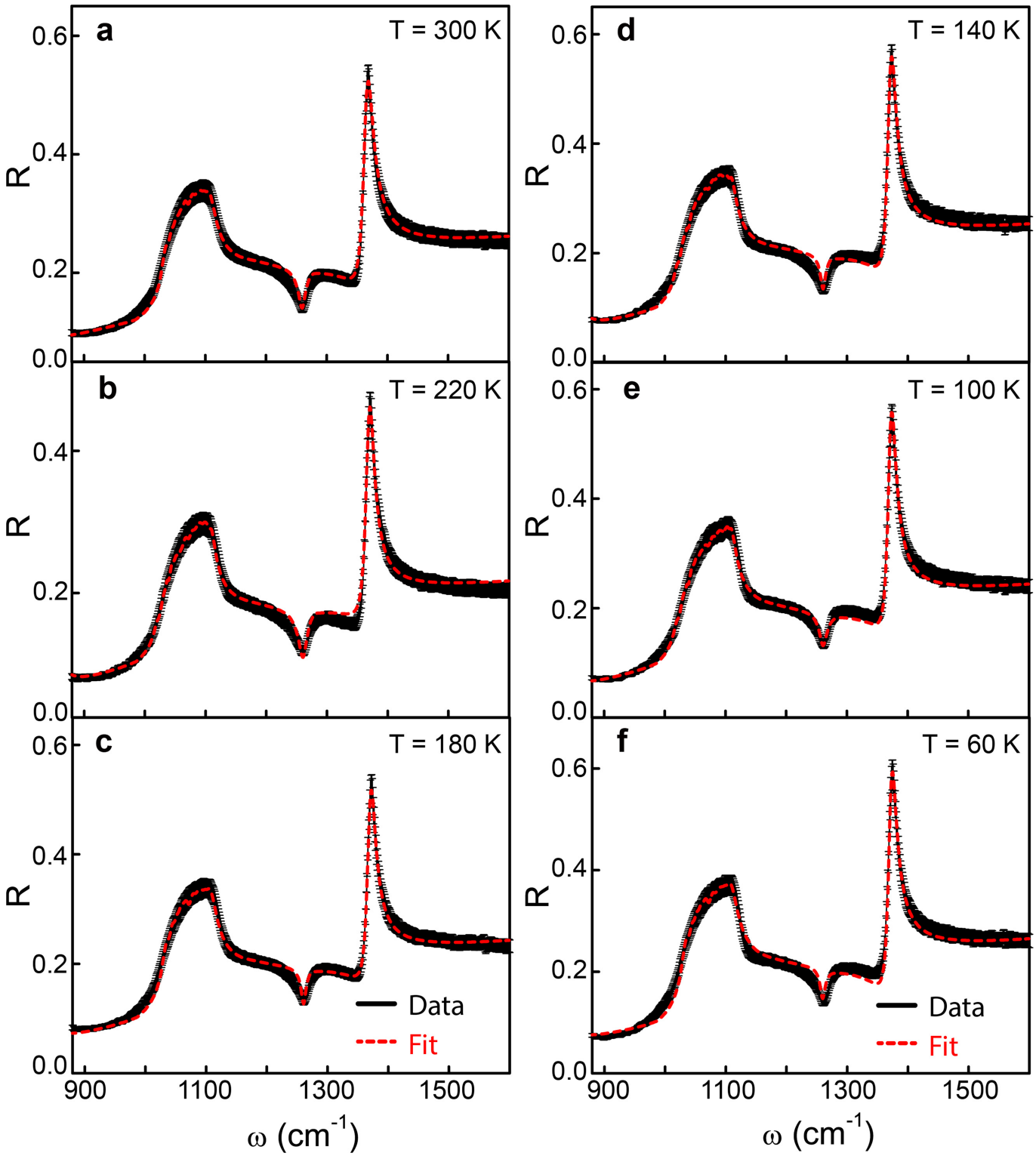
Extended Data Fig. 3 | Gate dependence of plasmon propagation at cryogenic temperature. a, Near-field amplitude *s* as a function of the gate voltage and the distance *L* from the Au launcher at a fixed frequency of $\omega = 886\ \mathrm{cm}^{-1}$. **b**, Illustration of the gate sweeping sequence. **c**, Line profiles (averaged over ~200 nm perpendicular to the propagation direction) of plasmonic interference fringes at different gate voltages.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



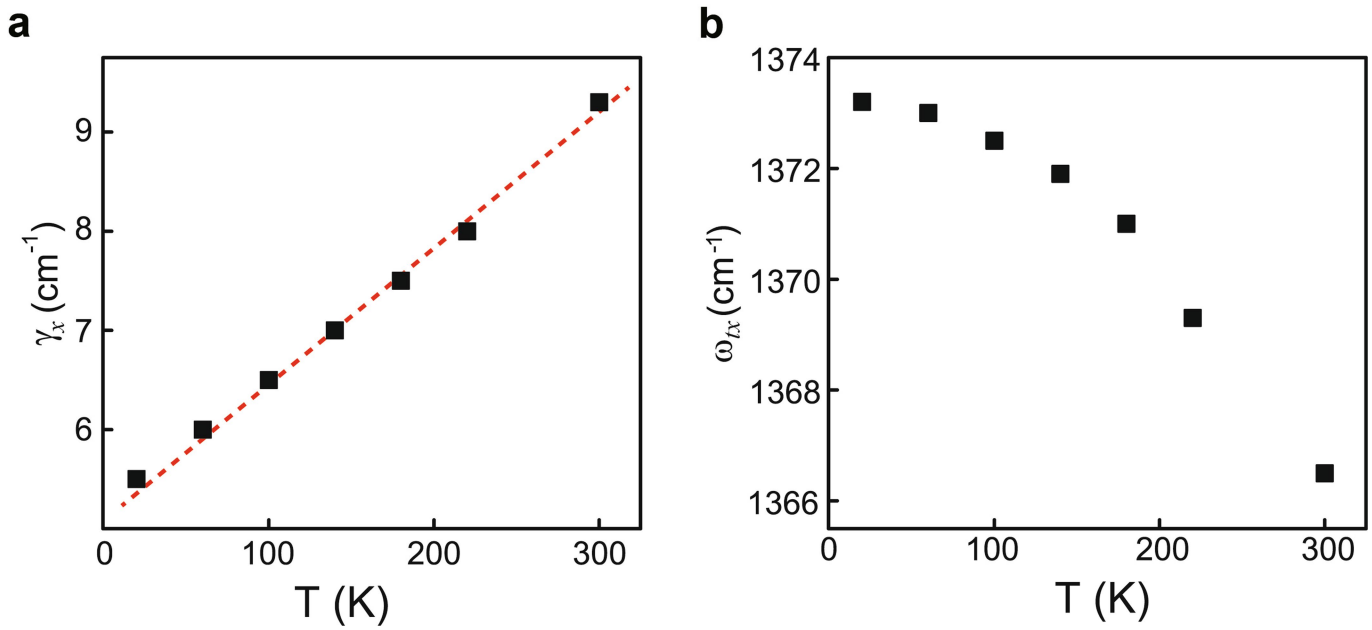
Extended Data Fig. 4 | Optical image of an hBN sample on a SiO_2/Si substrate. The dashed red square marks the approximate location of the region used in the reflectance measurements. The hBN thickness is 17 nm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



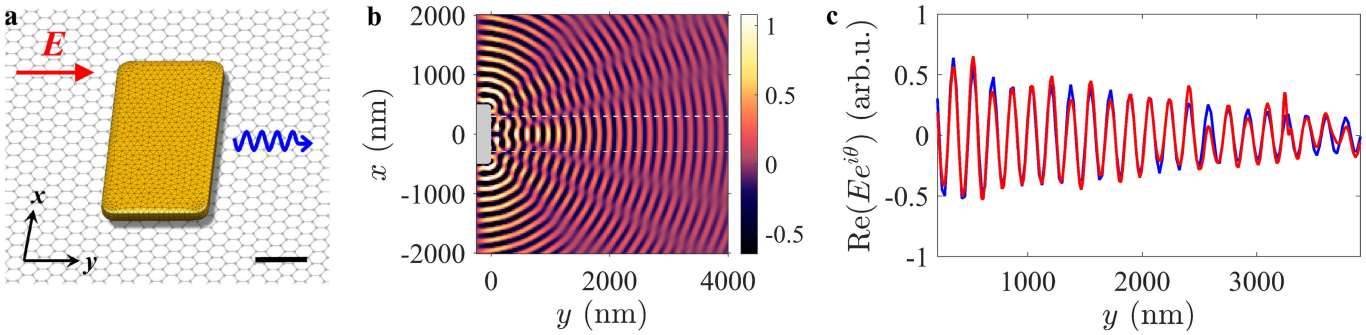
Extended Data Fig. 5 | Reflectance spectra of the hBN/SiO_2/Si structure at T=60–300 K. a–f, The black points are the experimental data and the red dashed lines are the fits. The sharp resonance at 1,370 cm^{-1} is due to the in-plane optical phonon of hBN, and the broad peak at around 1,070 cm^{-1} is due to the optical phonon of SiO_2.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



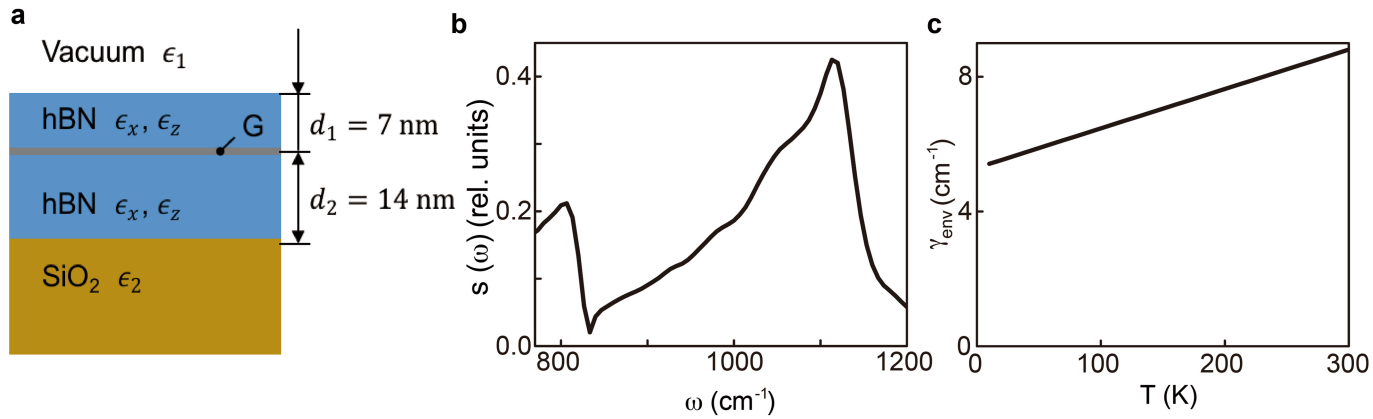
Extended Data Fig. 6 | The hBN in-plane phonon parameters versus the temperature. a, The fitted phonon linewidth γ_x. The squares are the data and the dashed line is a guide for the eye. **b**, The fitted frequency ω_{tx}.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



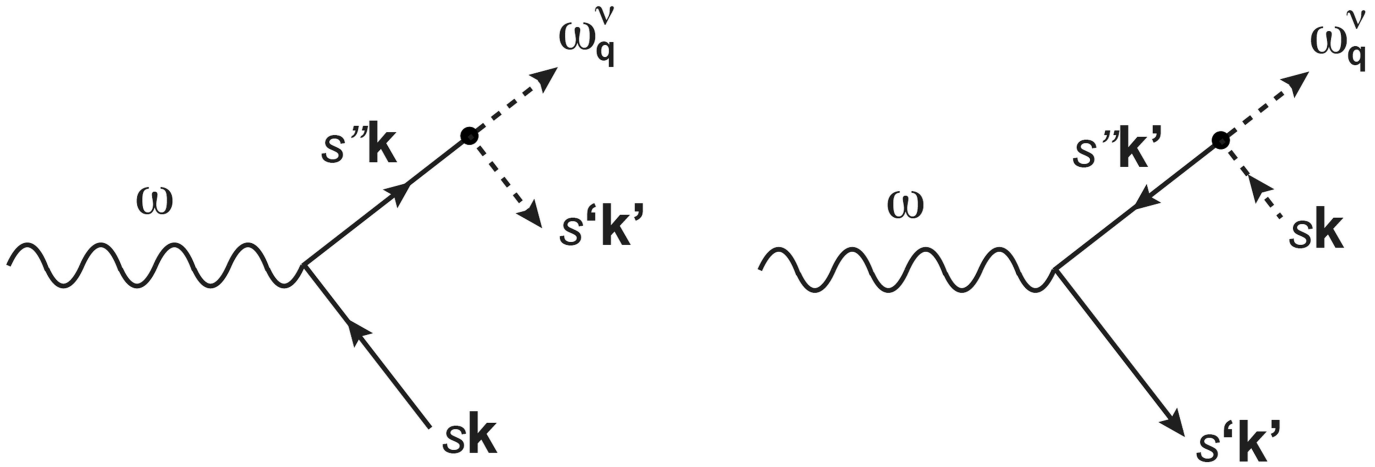
Extended Data Fig. 7 | Plasmonic oscillations in the vicinity of the Au launching pads. a, Illustration of the Au pad (gold, with the triangular mesh used in the simulation) on graphene (honeycomb lattice). The red arrow depicts the direction of the external field. The blue arrow symbolizes launched plasmons. Scale bar, 200 nm. **b**, An example of the simulated electric field distribution (E^z, imaginary part) just above the graphene layer. The grey rectangle represents the right half of the Au pad. The simulation parameters are $\lambda_p = 170$ nm and $Q = 130$. **c**, Comparison of the theoretical fit (blue) and the experimental data (red) for $V_g = 75$ V and $T = 60$ K. Both the theoretical and experimental traces are obtained by averaging multiple line cuts inside the 600-nm-wide strip indicated by the white dashed lines in **b**. The phase shift $\theta = 112°$ (see text) was used to align the oscillations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



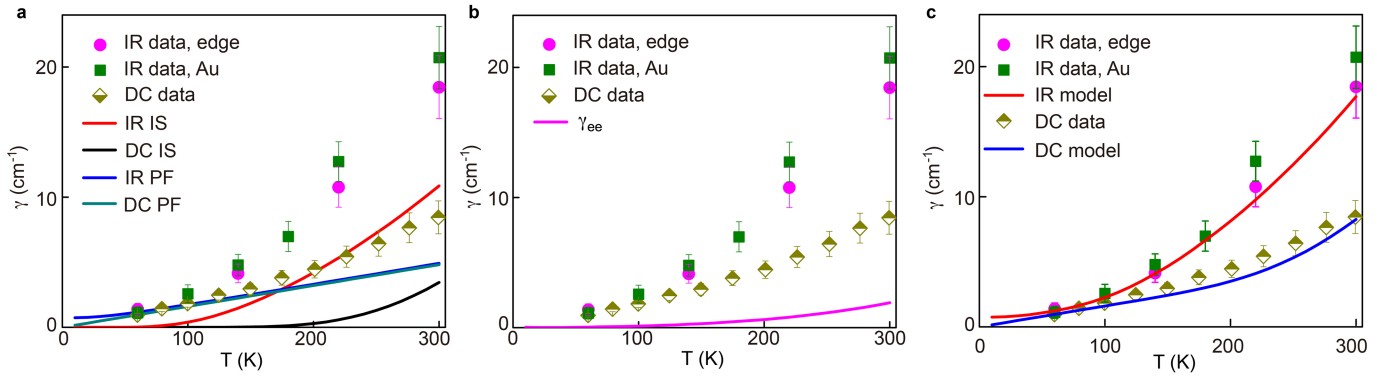
Extended Data Fig. 8 | Device structure and dielectric losses. **a**, Schematic of the device, showing the notations for the permittivities and thicknesses of the layers. **b**, Nano-FTIR spectrum $s(\omega)$ obtained with the hBN crystal taken away from the sample edges. **c**, The contribution γ_{env} of the dielectric environment to the plasmon linewidth as a function of temperature at a frequency of $\omega = 886\,\mathrm{cm}^{-1}$. The hBN c-axis damping constant is $\gamma_z(300\,\mathrm{K}) = 3.4\,\mathrm{cm}^{-1}$, which is consistent with previous

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



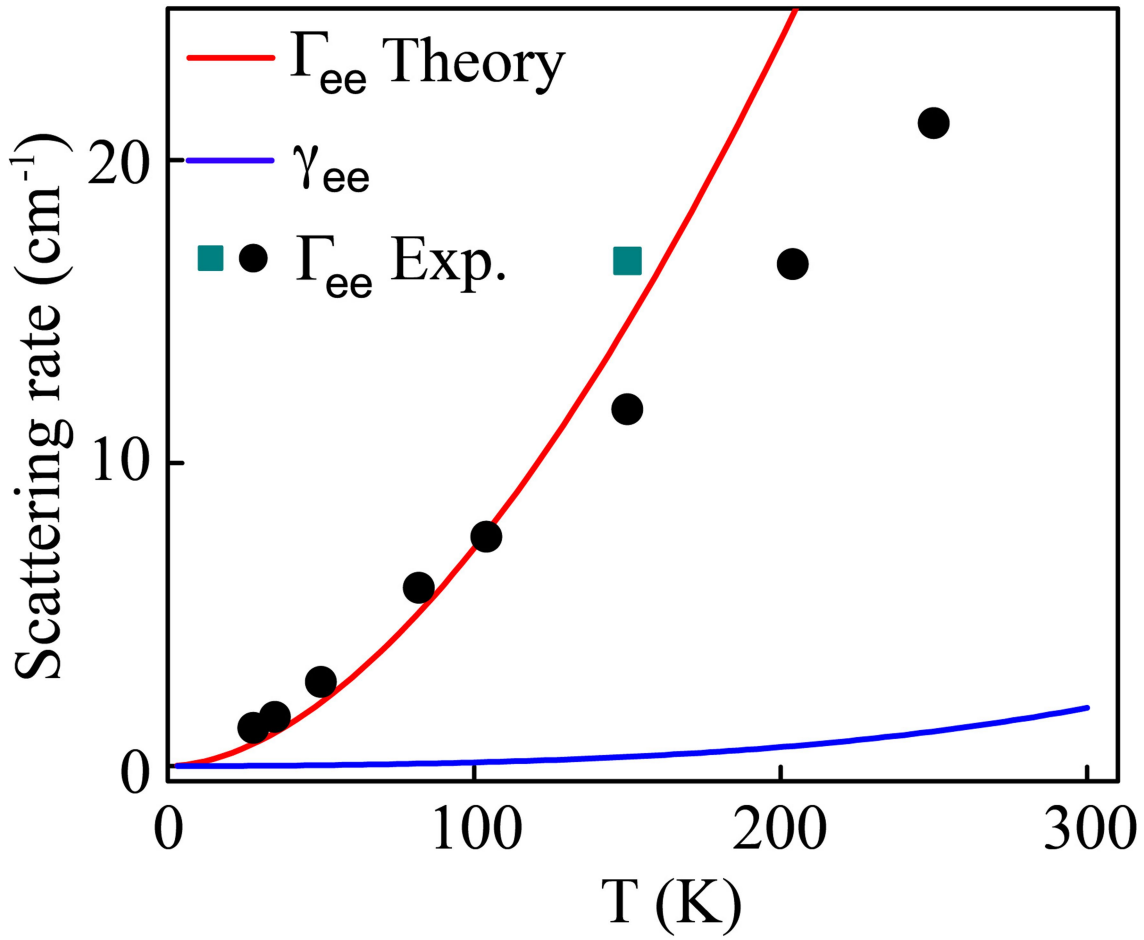
Extended Data Fig. 9 | Electron–phonon scattering processes. Diagrams of the scattering processes included in equation (12). The wavy, straight and dashed lines represent photons, electrons and phonons, respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Electron–phonon scattering rate as a function of temperature. a–c, Temperature dependence of the plasmonic scattering rate and of the d.c. scattering rate. Solid lines in **a** and **b** display the results of parameter-free modelling for electron–phonon scattering contributions (**a**) and electron–electron scattering contributions (**b**). Solid lines in **c** display the results of the sum in **a** and **b**, as discussed in the main text.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 11 | Plasmon damping rate due to electron–electron scattering as a function of temperature. The blue solid curve is the plasmon damping rate γ_{ee} due to electron–electron interactions, computed from equations (17) and (18) for a Fermi energy of $\varepsilon_F \equiv \mu(T=0) = 0.27\,\mathrm{eV}$. The red solid curve is the electron collision rate Γ_{ee} from equation (16). The squares and circles are Γ_{ee} values extracted from recent d.c. transport studies[23,24] at a different carrier density, $n \approx 10^{12}\,\mathrm{cm}^{-2}$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Phonon oscillator parameters for SiO_2

$\omega_{tj}\,(cm^{-1})$	$\omega_{pj}\,(cm^{-1})$	$\gamma_j\,(cm^{-1})$
1108.5	179.8	15.1
1104	219.7	12.6
1098	203.4	12.76
1092	330.8	19.6
1086	392.4	19
1077	355	13.76
1067	344	13.2
1059	386.9	16
1052	221	9.7
1045	302.8	13.2
1035	237.7	13
1030	37.68	16.9

The fitting parameters of equation (6) for $T = 300$ K.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Phonon oscillator parameters for hBN

Axis, *a*	$\varepsilon_{\infty a}$	$\omega_{ta}\,(\mathrm{cm}^{-1})$	$\omega_{pa}\,(\mathrm{cm}^{-1})$	$\gamma_a\,(\mathrm{cm}^{-1})$
t	4.90	Extended Data Fig. 6b	869	Extended Data Fig. 6a
z	2.95	760	346	Extended Data Fig. 8b

The single-Lorentzian fitting parameters of equation (6) for $T = 300\,\mathrm{K}$. See also Extended Data Figs. 6, 8.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0095-1

Addressing the minimum fleet problem in on-demand urban mobility

M. M. Vazifeh[1]*, P. Santi[1,2], G. Resta[2], S. H. Strogatz[3] & C. Ratti[1,4]

Information and communication technologies have opened the way to new solutions for urban mobility that provide better ways to match individuals with on-demand vehicles. However, a fundamental unsolved problem is how best to size and operate a fleet of vehicles, given a certain demand for personal mobility. Previous studies[1–5] either do not provide a scalable solution or require changes in human attitudes towards mobility. Here we provide a network-based solution to the following 'minimum fleet problem', given a collection of trips (specified by origin, destination and start time), of how to determine the minimum number of vehicles needed to serve all the trips without incurring any delay to the passengers. By introducing the notion of a 'vehicle-sharing network', we present an optimal computationally efficient solution to the problem, as well as a nearly optimal solution amenable to real-time implementation. We test both solutions on a dataset of 150 million taxi trips taken in the city of New York over one year[6]. The real-time implementation of the method with near-optimal service levels allows a 30 per cent reduction in fleet size compared to current taxi operation. Although constraints on driver availability and the existence of abnormal trip demands may lead to a relatively larger optimal value for the fleet size than that predicted here, the fleet size remains robust for a wide range of variations in historical trip demand. These predicted reductions in fleet size follow directly from a reorganization of taxi dispatching that could be implemented with a simple urban app; they do not assume ride sharing[7–9], nor require changes to regulations, business models, or human attitudes towards mobility to become effective. Our results could become even more relevant in the years ahead as fleets of networked, self-driving cars become commonplace[10–14].

Two trends—the rise of the autonomous and connected car, and the emergence of a 'sharing economy'[10,11] of transportation—seem poised to revolutionize the way personal mobility needs will be addressed in cities. The way current modes of transportation such as the private car, taxi or bus operate will be challenged and increasingly replaced by personalized, on-demand mobility systems operated by vehicle fleets, similar to what companies like Uber and Lyft offer. If such trends continue, they could lead to the displacement, or eventual disappearance, of jobs for bus and taxi drivers. Along with these possible societal costs, the transportation revolution could also offer immense benefits, including opportunities to resolve existing inefficiencies in individual urban mobility[12–14], thereby reducing traffic, whose carbon footprint currently accounts for about 23% of global greenhouse gas emissions[15,16].

To turn these opportunities into tangible environmental and societal benefits, autonomous and on-demand mobility systems need to be designed and optimized for efficiency, and integrated with carbon-efficient public transport. This requires the definition of models and algorithms for the evaluation of shared mobility systems that are both computationally efficient and accurate. The former property is mandated by the need to cope with hundreds of thousands (or sometimes millions) of trips routinely occurring in a large city. The latter property determines the relevance of the model results to the real world.

In what follows, we solve the 'minimum fleet problem' for the general case of on-demand mobility, and show that its solution for a specific case—taxi trips—could lead to breakthroughs in operational efficiency. To the best of our knowledge, no publicly available solution currently exists to address this minimum fleet-size problem at the urban scale for on-demand mobility in both private and public sectors. On the one hand, accurate methods based on mathematical programming (as traditionally used in the design of transportation systems[1–5,9]) can handle only a few thousand trips or vehicles at most, which is well below the hundreds of thousands or even millions of trips or vehicles routinely operating in large cities. On the other hand, city-scale studies[17] are obtained using a model of transportation based on aggregated mobility data and Euclidean spatial assumptions, and hence lack the resolution necessary to estimate the urban-scale benefits of vehicle sharing accurately.

We start from the notion of the shareability network introduced in ref. [7], which did not focus on the dispatching of vehicles. The type of shareability network introduced here is profoundly different from the type studied previously: it models the sharing of vehicles, whereas previous networks[7–9] modelled the sharing of rides. The main methodological contribution of this Letter is to show how this vehicle-sharing network can be translated into an exact formulation of the minimum fleet problem as a minimum path cover problem on directed graphs, thus establishing a connection to the rich applied mathematics and computer science field of graph algorithms. Besides revealing a structural property of vehicle-sharing networks, this connection allows the derivation of computationally efficient algorithms for optimal vehicle deployment and dispatching. Although optimally solving the minimum fleet size problem requires offline knowledge of daily mobility demand, in the following we also present a near-optimal, online version of the algorithm that can be executed in real time knowing only a small amount of the trip demand.

We are given a collection $\mathcal{T}$ of individual trips representing a portion of urban mobility demand during a certain time interval, such as a day. Each trip $T_i \in \mathcal{T}$ is defined as a tuple $(t_i^{\mathrm{p}},\ t_i^{\mathrm{d}},\ l_i^{\mathrm{p}},\ l_i^{\mathrm{d}})$ where t_i^{p} represents the desired pick-up time, l_i^{p} the pick-up location, t_i^{d} the drop-off time, and l_i^{d} the drop-off location. Here, the pick-up time means the earliest time t_i^{p} at which the passenger can be picked up at location l_i^{p}. The drop-off time means the estimated time of dropping off the passenger, calculated using a travel-time estimation model and assuming the passenger leaves the pick-up location at time t_i^{p}. In contrast to ref. [17], travel times here are computed using the actual road network, and using global positioning system (GPS)-based estimations derived from the taxi trip dataset that account for hourly variations in traffic, as in ref. [7]. If the set $\mathcal{T}$ is extracted from a real-world dataset (for example, taxi trips), the times t_i^{p} and t_i^{d} represent the actual times at which a passenger is picked up and dropped off, respectively.

The minimum fleet problem is formally defined as follows: 'find the minimum number of vehicles needed to serve all trips in $\mathcal{T}$, given that a vehicle is available at each l_i^{p} on or before t_i^{p}'. A service designed around this problem is ideal from a passenger's perspective, since a

[1]Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Istituto di Informatica e Telematica del CNR, Pisa, Italy. [3]Department of Mathematics, Cornell University, Ithaca, NY, USA. [4]Cornell Tech, New York City, NY, USA. [5]These authors contributed equally: M. M. Vazifeh, P. Santi. *e-mail: mmvazifeh@gmail.com

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

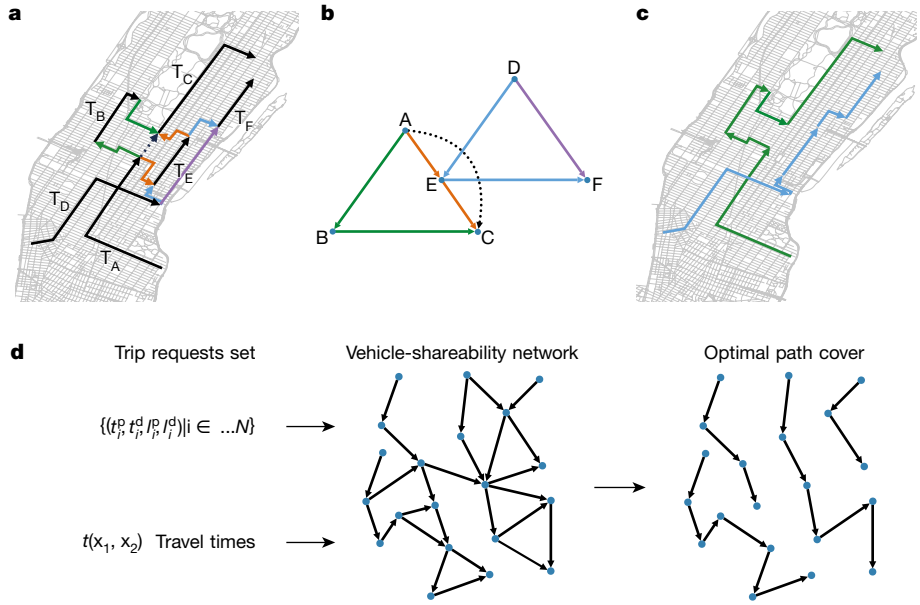


Fig. 1 | Constructing the vehicle-shareability network. a, Several trip requests are depicted on the map index as $T_A,\ldots, T_F$. The coloured path on the map represents different possibilities for vehicle dispatching. **b**, Colours are used in the vehicle-shareability network to specify how various dispatches can be represented using paths on this network. One of the two dispatching scenarios required only two vehicles whereas the other required three. **c**, The optimal vehicle-dispatching routes are represented on the map. **d**, To construct the vehicle-shareability network, trip set $\mathcal{T}$ and the travel times are taken as inputs. Two trips are connected by a directed edge if there is a large enough gap in time between the drop-off of the first and the pick-up point of the next trip to allow a single vehicle to travel between the two points before the pick-up time of the second trip starts, according to the travel-time information. Furthermore, the upper bound δ on the trip connection time must not be exceeded. The path-covering algorithm yields the path set that covers the entire node set, ensuring that all trips are served, while minimizing the number of paths (vehicles) in the solution. This is the optimal solution to the minimum fleet problem with parameter δ.

vehicle is guaranteed to be available at the desired location and time. On the other hand, the above problem formulation might entail substantial inefficiencies for the operator and the environment. Consider two consecutive trips T_A and T_B served by a single vehicle, and call the time needed to connect them the (trip) connection time, formally $t_{AB} = t_B^p - t_A^d$. If this time is very long, say, a few hours, it is trivially possible to connect trips that occur at distant locations or times. Hence, an excessively large connection time leads to inefficiencies for the operator (longer travelled distances, lower vehicle occupancy ratio) and the citizens (a lot of emissions and traffic just to connect trips). We therefore re-formulate the problem as follows: 'find the minimum number of vehicles needed to serve all trips in $\mathcal{T}$, under the assumptions that (1) a vehicle is available at each l_i^p on or before t_i^p and (2) the connection time is at most δ minutes', where the upper bound δ on the connection time is a problem parameter.

Figure 1 illustrates the construction of the vehicle-shareability network that enables the minimum fleet problem to be optimally solved with parameter δ. This is a directed network defined as $V = (N, E)$, where node $n_i \in N$ corresponds to trip $T_i \in \mathcal{T}$ and the directed edge $(n_i, n_j) \in E$ if and only if $(t_i^d + t_{ij}) \leq t_j^p$ (which accounts for assumption (1) above) and $t_j^p - t_i^d \leq \delta$ (which accounts for assumption (2) above). Here, t_{ij} represents the estimated travel time between l_i^d and l_j^p. The existence of a link in the network indicates that the two incident trips can be consecutively served by a single vehicle, and a path in V corresponds to a sequence of trips that can be served by a single vehicle—that is, a dispatch. Therefore, solving the minimum fleet problem is equivalent to finding the number of paths (vehicles) in the minimum path cover of V. The solution also gives the optimal dispatching strategy, that is, a sequence of trips to be served for each vehicle in the minimum fleet. The problem of finding the minimum path cover on general graphs is NP-hard, but it can be solved efficiently on directed acyclic graphs[18] using the Hopcroft–Karp algorithm for bipartite matching[19]. The acyclic nature of time guarantees that any vehicle shareability network is a directed acyclic graph, and the minimum fleet problem can be efficiently and optimally solved; see Methods for formal proofs.

We have tested our methodology on a dataset of over 150 million taxi trips performed in the city of New York in the year 2011. This dataset has been selected from a number of available datasets[8] because it is publicly available and, thanks to taxi statistics published by the New York Taxi and Limousine Commission[6], it is possible to compare our methodology directly with current taxi operation. The data have been sliced into daily datasets $\mathcal{T}_i$, each of which is an input to the minimum fleet size problem.

Next, we discuss how to set the parameter δ. When δ is decreased to 0, we approach a situation in which each trip is served by a dedicated vehicle: a solution with maximal vehicle utilization that is also optimal for traffic—under the assumption that vehicles materialize at the origin and dematerialize at the destination of the served trip—but incurring prohibitive costs for the mobility operator. On the other hand, when δ grows excessively, the fleet size is reduced, but the operational and traffic efficiency problems described previously occur. Thus, the setting of δ is an important design choice that is left to mobility operators, traffic authorities and policy makers. In this study, we set $\delta = 15$ min, as explained in Methods. The results of our method with different values of δ are reported in Methods (see Extended Data Fig. 1).

Figure 2 shows the daily number of vehicles needed to address the entire taxi demand in New York City using our approach. The minimum number of vehicles needed to serve trips is correlated with the number of daily trips (see Fig. 2a), with an overall R^2 value of 0.74. However, for the vast majority of days having between 300,000 and 550,000 trips (inset to Fig. 2a) this correlation becomes much weaker, with an R^2 value of only 0.18. Thus, trip density is a first determinant of fleet size, but trip spatiotemporal patterns are likely to play a large part as well. To investigate this issue further, we have analysed daily vehicle usage in the optimal solution.

The vehicle usage analysis reported in Methods shows that a fraction of vehicles, ranging between 5% and 10%, are highly underutilized and serve only around 1% of the trips, a lower utilization pattern that occurs especially during the weekend and is probably related to the extra nightly demand. The analysis also highlighted clear weekly patterns in

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

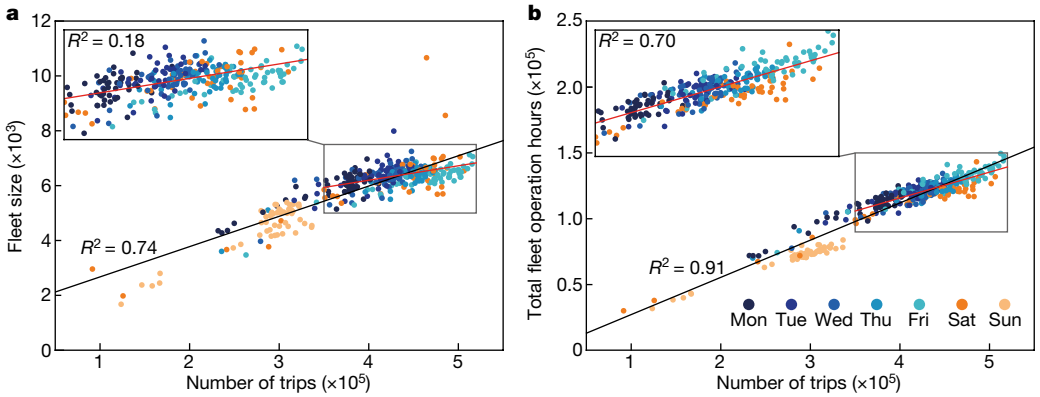


Fig. 2 | Minimum fleet-size analysis. a, The interplay between the number of daily trips and the minimum number of vehicles required to serve them. The colours of the dots correspond to different weekdays. Clustering of points with the same colour suggests weekly patterns, which are confirmed by the yearly analysis reported in Methods (see Extended Data Fig. 2). The plot shows a moderate correlation between the two quantities. However, when focusing on days with a number of trips ranging from 350,000 to 550,000 (inset), the correlation becomes much weaker, as the fleet size remains within a narrow window of around 6,000 vehicles. **b**, The correlation between the number of daily trips and the total fleet operation hours. The latter is defined as the total time of operation, estimated by summing over the operation times for each vehicle in the minimum fleet obtained over a day. The operation time for each vehicle is defined as the total time a vehicle is operating on the road to serve the trips. The total fleet operation hours have a much stronger correlation with the number of trips than with the number of vehicles.

vehicle use, consistent with the relatively stable vehicle fleet size across the year. This observed stability can be explained by a simple model for vehicle trip assignment, and is fundamental for mobility operators: it indicates that investment in acquiring an optimal number of vehicles for operation gives consistent yearly returns. The dip in vehicle fleet size occurring at weekends hints also at an opportunity to perform routine vehicle maintenance on a weekly basis.

A better scaling law relating vehicle fleet size to the daily number of trips can be obtained by defining a metric for fleet sizing that incorporates how long a vehicle is used during a day. We define a 'full-time equivalent' vehicle as a vehicle continuously operating 24 h a day. (In the case of human-driven vehicles, we can think of having the vehicle operated in three 8-h shifts, for instance.) Figure 2b shows that the scaling law relating the number of daily trips with full-time equivalent vehicles is more accurate than the previous one, with the coefficient of determination R^2 value increased from 0.74 to 0.91, and from 0.18 to 0.70 for the trip-intense days reported in the inset.

Figure 3 shows the efficiency breakthrough provided by network-based optimization: when compared to current taxi operation in New York City, the number of circulating taxis can be reduced by an impressive 40%, and kept fairly constant through the day. This improvement is all the more noticeable considering that it is achieved without imposing any delay on customers, nor sharing of rides as in refs. [7,9]. That fleet size can be reduced by as much as 40% without the use of ride sharing and with no delay for passengers has, to the best of our knowledge, not been reported in the literature before, and it is one of the main results of this paper.

The 40% fleet reduction reported above refers to the model with full knowledge of daily trip demand. If only a portion of trip demand is known, as in current on-demand mobility services where trip requests are collected in real time, we can still achieve near-optimal performance

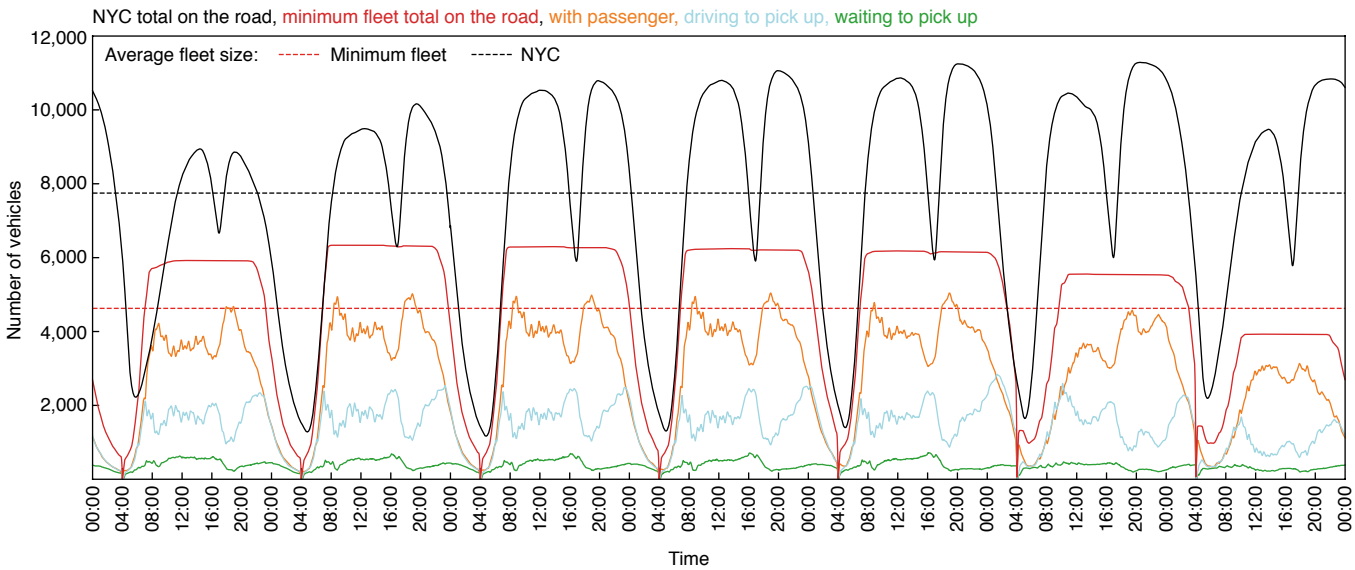


Fig. 3 | Fleet efficiency comparison. Comparison between actual number of New York City (NYC) taxis on the road[6] (black curve), versus the minimum number of vehicles as computed by our optimal approach (red curve). For the optimized approach, a breakdown of vehicles into 'with passenger', 'waiting to pick up' and 'driving to pick up' is also reported. The curves are computed by averaging data across one year. As shown by the dotted lines corresponding to average deployment, optimized dispatching brings a reduction in number of circulating taxis from 7,748 down to 4,627, a 40% reduction. With online operation ready to implement using a smartphone app, our method provides a near-optimal 30% reduction in the number of circulating taxis (see Fig. 4).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

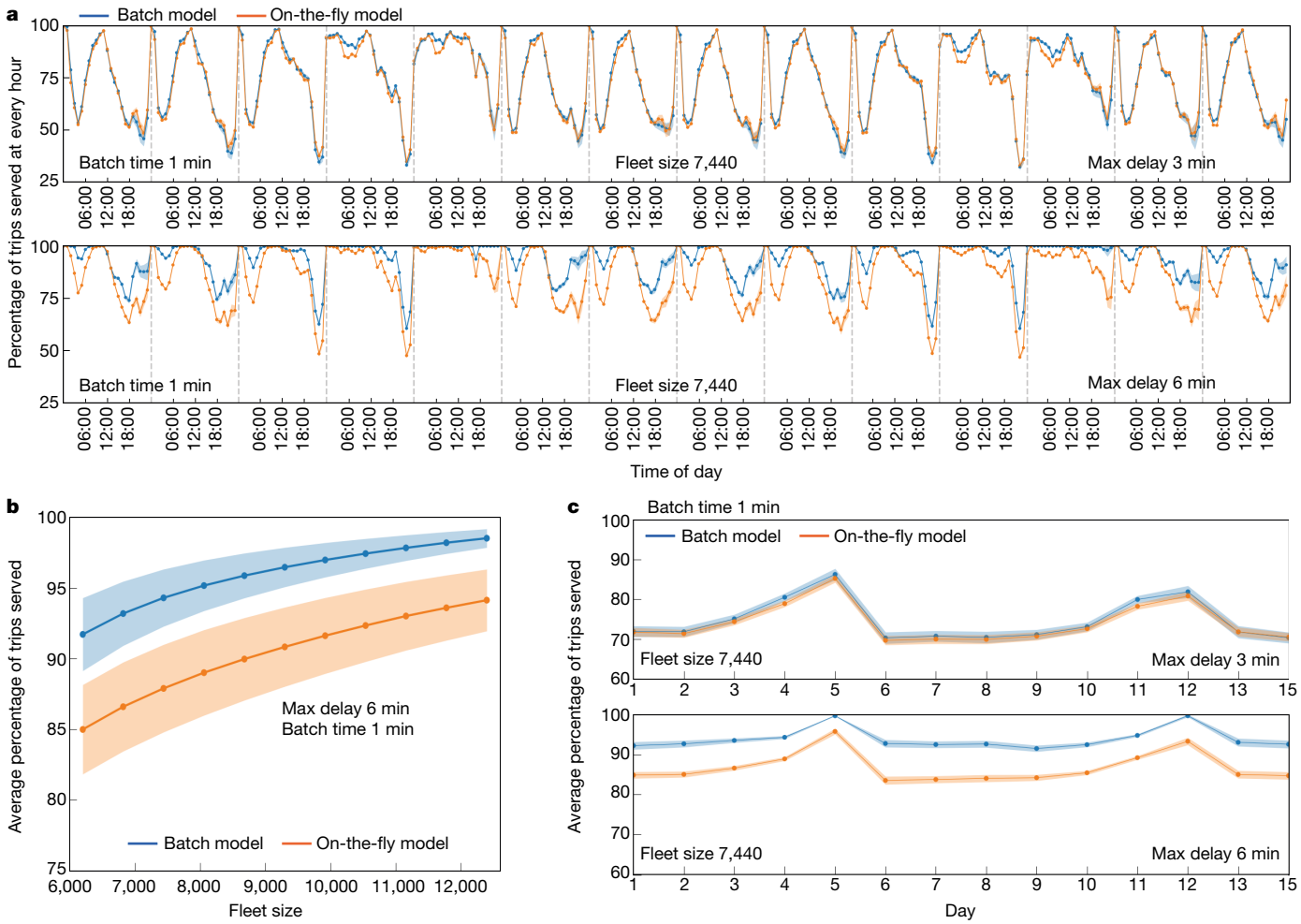


Fig. 4 | Performance of the network-based online vehicle-dispatching model. a, Two-panel plot comparing the percentage of trips served within a certain maximum delay max(Δt) based on the batch-based optimized dispatching (blue line) versus sequential dispatching, as described in the Methods. The panel at the top represents the results for max(Δt) = 3 min and the one at the bottom is for max(Δt) = 6 min. The use of network-based dispatching (blue line) provides a substantial improvement in the percentage of successfully served trips with respect to sequential dispatching (orange line). The fleet size is set to $N = N_{\min}x$, where $N_{\min}$ is the average minimum fleet size obtained from running the algorithm with daily trip knowledge based on the historical data for the 15 days considered in the evaluation. The fleet size inflation ratio x is set to 1.2, the maximum passenger waiting time is set to max(Δt) = 3 min or 6 min, and the batching time in batch-based dispatching is set to 1 min. **b**, The average percentage of trips successfully served within a 6-min delay is above 90% for a fleet of 6,200 vehicles. Similar performance can only be achieved with sequential dispatching when the fleet size is more than 10,000. **c**, Plots showing daily averages of the total percentage of trips in a day successfully served within the specified delay for the same days as in **a**.

with the online version of the algorithm reported in the Methods. This version collects trip requests for a short time, for example, one minute, and locally optimizes vehicle dispatching based on this limited knowledge. Figure 4 shows that, with a 30% fleet reduction and using the online version of the algorithm, more than 90% of the trip requests can be successfully served, hitting a performance very close to the 40% fleet reduction possible when the entire daily demand is known beforehand.

Our approach assumes that trip requests and vehicle-dispatching decisions are centralized, a model that is radically different from current taxi operation and similar to the one used by online mobility operators. Therefore, the benefits of optimized operation reported in Fig. 3 can be interpreted as being implied by the transition from a fully distributed operation, where the deployment strategy is based on individual driver decisions, to a centralized operation, where dispatching decisions are globally optimized. To some extent, our results can then be seen as a quantification of the well known game-theory notion of the 'price of anarchy'[20] in urban taxi operation. Taking a mobility market perspective, this is a transition from a regulated mobility market with numerous micro-operators (down to the level of the single taxi driver), to a monopolistic market with a single mobility operator with centralized operation. Although optimal from the vehicle operation and environmental viewpoint, a monopolistic market is however highly undesirable for many other reasons, most importantly, lack of competition with consequent higher prices for customers. An additional analysis reported in Methods shows that most of the efficiency benefits of centralized vehicle operation are still possible in an oligopolistic market.

Although the characterization of minimum fleet size reported here is fully representative of an autonomous driving scenario where human operation of vehicles is not necessary, constraints on driver availability and maximum operating hours, shift operation and so on might produce relatively larger values of the minimum fleet requirement than those predicted here. Extending the concept of the vehicle-sharing network to incorporate driver constraints is possible and is left for further analysis.

Broader effects on traffic are foreseen if our methodology is to be used for optimizing urban 'on-demand' mobility services more in general, especially in a future of autonomous vehicles. However, it is well known that an improvement in mobility efficiency is sometimes linked with an increase in demand which, in turn, could reduce the amount of traffic reductions. Evaluating this 'second-order' effect of optimized fleet operation on urban traffic requires coupling a micro-level traffic simulation, agent-based passenger models and our network-based methodology, a challenging task which we leave to future work.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Finally, we observe that, while applied here to taxi trips as a case study, the proposed methodology for optimal vehicle fleet sizing and dispatching is general and can be applied to model any type of point-to-point mobility. However, the approach presented here focuses on optimizing and dispatching a single fleet of vehicles. Optimization across different fleets and transportation modes is possible by extending our approach to consider multiple coexisting fleets of various types to serve the mobility demand. With the approaching advent of autonomous mobility and the forecast increase in sharing cars (or other autonomous vehicles, such as flying drones), the problem of how to optimize and orchestrate multiple autonomous fleets will come to the forefront, and might be addressed using the scalable and accurate analytical tools presented here for optimal solution of the 'minimum fleet' problem.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0095-1.

Received: 19 July 2017; Accepted: 21 February 2018;
Published online 23 May 2018.

1. Berbeglia, G., Cordeau, J. F. & Laporte, G. Dynamic pick-up and delivery problems. *Eur. J. Oper. Res.* **202**, 8–15 (2010).
2. Laporte, G. The vehicle routing problem: an overview of exact and approximate algorithms. *Eur. J. Oper. Res.* **59**, 345–358 (1992).
3. Baker, B. M. & Ayechew, M. A. A genetic algorithm for the vehicle routing problem. *Comput. Oper. Res.* **30**, 787–800 (2003).
4. Yang, J., Jaillet, P. & Mahmassani, H. Real-time multivehicle truckload pick-up and delivery problems. *Transport. Sci.* **38**, 135–148 (2004).
5. Clare, G. L. & Richards, A. G. Optimization of taxiway routing and runway scheduling. *Proc. IEEE Intell. Transport. Syst.* **12**(4), 1000–1013 (2011).
6. Bloomberg, M. R. & Yassky, D. *New York City Taxicab Factbook* http://www.nyc.gov/html/tlc/downloads/pdf/2014_tlc_factbook.pdf (New York City Taxi and Limousine Commission, 2014).
7. Santi, P. et al. Quantifying the benefits of vehicle pooling with shareability networks. *Proc. Natl Acad. Sci. USA* **111**, 13290–13294 (2014).
8. Tachet, R. et al. Scaling law of urban ride sharing. *Sci. Rep.* **7**, 42868 (2017).
9. Alonso-Mora, J., Samaranayake, S., Wallar, A., Frazzoli, E. & Rus, D. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proc. Natl Acad. Sci. USA* **114**, 462–467 (2017).
10. Rosenberg, T. It's not just nice to share, it's the future. *NY Times Opinionator* https://perma.cc/89YT-VHVF (2013).
11. Sundararajan, A. From Zipcar to the sharing economy. *Harvard Business Review* https://hbr.org/2013/01/from-zipcar-to-the-sharing-eco (2013).
12. Mitchell, W. J., Borroni-Bird, C. E. & Burns, L. D. *Reinventing the Automobile: Personal Urban Mobility for the 21st Century* (MIT Press, Cambridge, 2010).
13. Botsman, R. & Rogers, R. *What's Mine is Yours: The Rise of Collaborative Consumption* (HarperCollins, New York, 2010).
14. Handke, V. & Jonuschat, H. *Flexible Ridesharing* (Springer, Berlin, 2013).
15. United Nations Environment Programme *UNEP 2010 Annual Report*; https://www.unenvironment.org/resources/annual-report/unep-2010-annual-report (UNEP, Nairobi, 2011).
16. Barth, M. & Boriboonsomsin, K. Real-world carbon dioxide impacts of traffic congestion. *Transp. Res. Rec.* **2058**, 163–171 (2011).
17. Spieser, K. et al. in *Road Vehicle Automation* (eds Meyer, G. & Sven Beiker, S.) 229–245 (Springer, 2014).
18. Boesch, F. T. & Gimpel, J. F. Covering the points of a digraph with point-disjoint paths and its application to code optimization. *J. Assoc. Comput. Mach.* **24**, 192–198 (1977).
19. Hopcroft, J. & Karp, R. An $n^{(5)/(2)}$ algorithm for maximum matching in bipartite graphs. *SIAM J. Comput.* **2**, 225–231 (1973).
20. Roughgarden, T. *Selfish Routing and the Price of Anarchy* (MIT Press, Cambridge, 2005).

Acknowledgements P.S., M.M.V. and C.R. thank all sponsors and partners of the MIT Senseable City Laboratory including Allianz, the Amsterdam Institute for Advanced Metropolitan Solutions, the Fraunhofer Institute, Kuwait-MIT Center for Natural Resources and the Environment, Singapore-MIT Alliance for Research and Technology (SMART) and all the members of the Consortium. The research of S.H.S. was supported by USA NSF grant numbers DMS-1513179 and CCF-1522054.

Reviewer information *Nature* thanks G. Casey, V. Knoop and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions P.S. defined the problem, designed the solution and algorithms, and contributed to the analysis and paper writing. M.M.V. designed and performed the analysis, developed models and simulations and wrote the paper. G.R. contributed to the algorithm design, implemented the algorithms and helped with the analysis. S.H.S. contributed to writing. C.R. supervised the research and contributed to writing.

Competing interests The authors declare no competing interests.

Additional information
Extended data. is available for this paper at https://doi.org/10.1038/s41586-018-0095-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0095-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.M.V.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Trip data. The dataset used in this work consists of more than 150 million trips with passengers of all 13,586 taxicabs in New York City during the calendar year of 2011. The dataset contains a number of fields from which we use the following: origin time, origin longitude, origin latitude, destination longitude and destination latitude. The measurement precision of times is in seconds; location information has been collected by the data provider via GPS location tracking technology. Out of our control are possible biases due to urban canyons (that is, streets with high-rise buildings on both sides), which might have slightly distorted the GPS locations during the collection process. All individual-level identifications are given in anonymized form; origin and destination values refer to the origins and destinations of trips, respectively.

Map matching. Similar to the preprocessing done in ref. [7], we used data from www.openstreetmap.org to create the street network of Manhattan. As described in previous work[7], we used a filtering method on the streets of Manhattan to select only the following road classes: primary, secondary, tertiary, residential, unclassified, road and living street. We left out several other classes deliberately. These include footpaths, trunks, links or service roads, as they are unlikely to contain delivery or pick-up locations. We extracted the street intersections to build a network in which nodes are intersections and directed links are roads connecting those intersections (we use directed links because a non-negligible fraction of streets in Manhattan are one-way). The extracted network of street intersections was then manually cleaned for obvious inconsistencies or redundancies (such as duplicate intersection points at the same geographic positions), in the end containing 4,091 nodes and 9,452 directed links. This network was used to map match the GPS locations from the trip dataset. We matched only GPS locations (both trip origins and destinations) that are within 100 m of at least one node in the street intersection network, which is the case for the majority of trips, and discarded the remaining ones. After the preprocessing and filtering phase, more than 147 million trips remain to be used in the next phases of our analysis.

Travel-time computation. Travel-time information is a key part of building vehicle-shareability networks. The knowledge of estimated travel times is based on a heuristic method developed and used in ref. [7]. This method uses pick-up and drop-off times of a historical trip dataset and computes the travel times between arbitrary origins and destinations on the road map.

In the following we briefly describe the core idea of this method. A detailed description can be found in the supplementary information of ref. [7].

Each street segment belongs to the set, $S=\{S_1,\ldots,S_h\}$, of all road segments connecting any pair of adjacent intersections in the road map. Given a set of k historical trips $\mathcal{T}=\{T_1,\ldots,T_k\}$, the problem of travel-time computation is estimating the travel time t_i^e for each street segment $S_i\in S$ in such a way that the average relative error (computed across all trips) between the actual travel time t_i and the estimated travel time t_i^e for trip T_i computed starting from the x_i (compound with a routeing algorithm) is minimized. Once error-minimizing travel times for each street in S are determined, the travel time between any two intersections i and j can be computed starting from the t_i values, using a routeing algorithm that minimizes the travel time between any two intersections.

The following steps are involved in the process of travel-time computation. First, we partition the trip set in time-sliced subsets $\mathcal{T}_1,\ldots,\mathcal{T}_{24}$ where subset $\mathcal{T}_i$ contains all trips whose starting time is in hour i of the day. If desired, finer partitioning (for example, per hour and weekday, per hour and weekday and month, and so on) is possible. The travel-time estimation process can be performed independently on each of the time-sliced trip subsets. We define $\mathcal{T}_i^{sq}$ as the subset of trips with origin x_s and destination x_q in which x_s contains the (latitude, longitude) coordinates of the sth intersection after the trips are matched. A small fraction of trips are filtered to remove 'loop' trips (that is, trips with the same origin and destination), as well as excessively short or long trips. After a step in which initial routes are computed using a pre-selected initial speed v_{int} (the same for all streets) as described in ref. [7], a second trip-filtering step is performed, in which excessively fast and slow trips are removed from the travel-time estimation process. 97% of trips remain after this filtering. The travel-time estimations obtained using this method are reasonable, with a relatively lower average speed of around 5.5 m s^{-1} estimated during rush hours (between 8 am and 3 pm), and peaks at around 8.5 m s^{-1} at midnight.

Node-disjoint path cover. In the following we provide a set of definitions and present relevant theorems with their proofs to systematically formulate the problem of reducing the fleet size as a path-covering problem on a vehicle-shareability network.

Given a directed network $V=(N,E)$, a path P in V is a sequence of edges $\{e_1=(n_1^1,n_1^2),\ldots,e_k=(n_k^1,n_k^2)\}\in E$ such that $n_i^2=n_{i+1}^1$, for each $i=1,\ldots,k-1$. The set of nodes in path P is defined as $N(P)=\bigcup_{i=1,k}\{n_i^1\}$. The length of a path P is the number of edges k that form it.

Definition 1 (Path cover). Given a directed network $V=(N,E)$, a node-disjoint path cover of V is a collection of paths $\{P_1,\ldots,P_h\}$ such that $\cup_{i=1,h}N(P_i)=N$ and $N(P_i)\cap N(P_j)=\varnothing$ for any $i\neq j$. The size of the cover is the number of paths h of which it is formed.

We note that, under the conventional assumption that zero-length paths corresponding to single nodes are allowed, a node-disjoint path cover always exists. In the following, to simplify presentation we drop the term 'node-disjoint' and use 'path cover' to refer to a 'node-disjoint path cover' as defined in 'Definition 1 (Path cover)' above.

Theorem 1. Let $\mathcal{C}=\{P_1,\ldots,P_h\}$ be a path cover of the vehicle-shareability network $V=(N,E)$. Then, all the trips in $\mathcal{T}$ can be served by h vehicles.

Proof. Consider a path $P=\{e_1=(n_1^1,n_1^2),\ \ldots,\ e_k=(n_k^1,n_k^2)\}$ in the vehicle-shareability network V. By definition of shareability network, the trips corresponding to n_1^1 and n_1^2 (call them T_1 and T_2) can be served by a single vehicle. Furthermore, the vehicle performing trip T_1 is guaranteed to arrive at the pick-up location of T_2 within time t_2^p; that is, vehicle sharing does not impose any delay on the starting time of the second trip. Also, the upper bound δ on the trip connection time is not violated by the definition of shareability network. Hence, the vehicle that serves T_1 and T_2 can be used to serve trip T_3 corresponding to node n_2^2 in V, since the starting time of trip T_2 is not changed as a result of sharing, implying that the condition ensuring shareability of T_3 and T_2 is still fulfilled. By iterating the argument across all nodes in $N(P)$, we can conclude that all trips whose corresponding nodes are in $N(P)$ can be served by a single vehicle. Thus, if a path cover of size h exists, we can conclude that all trips in $\mathcal{T}$ can be served by h vehicles.

Corollary 1. The minimum number of vehicles needed to serve the trips in $\mathcal{T}$ equals the size of the minimum path cover of the vehicle shareability network $V=(N,E)$.

Finding the size of the minimum path cover of an arbitrary directed network is NP-hard[18] and hence computationally infeasible for large graphs. However, the optimal solution can be found in polynomial time if the network is acyclic, meaning that there is no directed path in the network forming a closed loop.

Definition 2 (Directed acyclic network). A directed network $V=(N,E)$ is acyclic if it has no directed cycles, that is, it does not contains directed paths starting at some vertex $n\in N$ and eventually returning to n again.

Any vehicle-shareability network as defined above is a directed acyclic network. To see how the acyclic character arises one can use proof by contradiction. Assume a cyclic path exists in V. For simplicity, assume the path has minimal length of 2. Let $P=\{(n_1,n_2),(n_2,n_1)\}$ be a cyclic path, and let T_1 and T_2 be the trips corresponding to n_1 and n_2, respectively. By the definition of vehicle-shareability network, we have the following sequence of inequalities:

$$t_1^d\le t_1^d+t_{12}\le t_2^p<t_2^d\le t_2^d+t_{21}\le t_1^p$$

which is a contradiction since $t_1^d>t_1^p$. Hence, no cyclic path of length 2 can exist in V. The proof follows by straightforwardly extending the above sequence of inequalities to cyclic paths of arbitrary length. This implies that the minimum number of vehicles needed to perform a set $\mathcal{T}$ of trips can be computed in polynomial time. More specifically, it is shown that for directed acyclic networks the problem of finding the path cover of minimum size is equivalent to the well known maximum matching problem on bipartite graphs, which can be solved in time $O(|E|(|N|)^{1/2})$ using the Hopcroft–Karp algorithm.

Online model. The results shown so far compute the minimum infrastructure on the basis of the knowledge of the entire shareability network for the day considered. This is analogous to the Oracle model as defined in ref. [7], and is consistent with a scenario in which trip requests are issued in advance (for example, through a reservation system). To investigate to what extent the above described benefits extend to systems where trip requests are issued in real time (such as Uber and Lyft), we repeat the analysis in the so-called online model. In the online model, we have a number of vehicles available for serving trips, which is defined as $N=N_{\min}x$, where $N_{\min}$ is the minimum fleet size for the day of reference as computed by the oracle model, and $x>1$ is an inflating factor. We then start serving trip requests with the available vehicles, whose initial position is determined through a warm-up phase in which a number of trip requests from the previous day (not accounted to compute the results) are served. To compare online models, two possible strategies are used to dispatch vehicles and serve trip requests, as follows.

On-the-fly. Trip requests are served sequentially; when a new trip request is issued, the dispatched vehicle is chosen as the first available vehicle that minimizes passenger waiting time.

Batch. Trip requests are collected for time $\delta=1$ min and processed in batches. When a batch is processed, a maximum matching is computed to maximize the number of requests that can be successfully served (that is, served within max-$(\Delta t)=6$ min); vehicles are then dispatched on the basis of the result of the maximum matching algorithm, as explained in the following. At each given minute the trip requests information and the locations of the available vehicles are compiled to construct a weighted bipartite graph. The edge weight on a pair of vehicle-trip node represents the pick-up delay a passenger associated with the trip node would

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

experience in case the vehicle associated with the vehicle node is chosen to serve the passenger. After constructing this weighted bipartite network, the maximum matching algorithm can be used to find a subset of edges covering the maximum number of trip nodes served within the tolerable delay, $\max(\Delta t)$.

Figure 4 shows the success rate of the two dispatching algorithms for a period of 15 consecutive days, for $x = 1.2$, serving the trips within a certain tolerable delay. As seen in this figure, the batch method (blue lines) provides a success rate which is consistently above 92%, and is much higher than what is achieved by the sequential on-the-fly method for $\max(\Delta t) = 6$ min. As reported in Extended Data Table 1, the running times of the online version of the method are below 200 ms in the worst-case scenarios on a standard Linux machine, indicating the feasibility of the proposed approach for real-time optimization.

The warm-up phase used in the above-mentioned online optimizations consists of first deploying each vehicle at a random intersection, then running the batch optimization algorithm as described above on the 2 h of historical trip requests that precede the period of interest. The shaded regions in Fig. 4a and c and in Extended Data Fig. 3 represent the variation in the percentage of the trip served as obtained by running the real-time optimization for each day multiple times, each time reinitializing the warm-up phase with a distinct random initial deployment of the fleet. The variations are quite small, showing that within 2 h the system's spatiotemporal distribution does not depend noticeably on the initial deployment.

Limiting node connectivity via trip connection time. We defined the vehicle-shareability network in such a way that nodes that represent individual trips are connected only if it is feasible for a vehicle to serve those trips one after the other without introducing any delay in their pick-up and drop-off times. Checking whether two trips satisfy such criteria requires knowledge of travel times in the city, which is estimated using the method described previously. Since this network definition puts no constraints on the connectivity apart from the feasibility of consecutively serving trips, the number of network links grows quickly with the increase in the number of trips. This is because trips separated by a large enough time gap between their drop-off and pick-up times can always be served by the same vehicle although they may be spatially far from each other. This leads to a very high connectivity in the vehicle-shareability network because most pairs of trips separated by enough time can satisfy this connectivity condition. To limit the number of edges in the network, and to make sure that the vehicles do not operate without any passenger onboard for too long leading to underutilization and an increase in the void ratio (the fraction of time vehicles operate without a passenger), we introduce an upper bound on the connection time between the trips. The connection time is defined as the time a vehicle operates without a passenger between the consecutive trips.

The first issue to address is how to set the bound δ on the trip connection time, which is a parameter that can be used to trade off fleet size against vehicle and traffic efficiency. On the one hand, when δ is decreased to 0 we approach a situation in which each trip is served by a dedicated vehicle: a solution with maximum vehicle utilization that is also optimal for traffic (if we assume that vehicles somehow appear at the origin and disappear at the destination of the trip they serve), but incurring prohibitive costs for the mobility operator. On the other hand, when δ grows excessively the fleet size is reduced, but this is at the expense of a decrease in the operational and traffic efficiency because some vehicles may be on the road for long times without any passenger on board between serving the trips. Thus, how to set δ is an important design choice, which should be left in the hands of mobility operators, traffic authorities and policy makers.

Extended Data Fig. 1 shows how we come up with a reasonable setting for δ. The plot reports both the minimum fleet size as well as the average fraction of time a vehicle spends connecting consecutive trips (the void ratio) in seconds, for increasing values of δ. As expected, the former quantity decreases with δ, while the latter increases. For values of δ larger than 15 min, however, the vehicle fleet size decreases only marginally, whereas the void ratio still increases. For this reason, for the results reported in the main text we have set $\delta = 15$ min. For reference, the right panel of Extended Data Fig. 1 reports the yearly analysis of minimum fleet size—similar to what is reported in Extended Data Fig. 2—for $\delta = 10$ min and 20 min.

Vehicle utilization. A better understanding of the efficiency of the network-based vehicle-trip assignment requires a closer look into the patterns of the utilization of the individual vehicles in the minimum fleet. The overall time each vehicle spends during its operation in a day consists of travelling with a passenger on board, without any passenger and on the way to pick up the next one, or waiting at the pick-up location of a new passenger. Ultimately, the goal in an efficient vehicle–trip assignment is to maximize overall utilization while minimizing the operation costs. This is achieved for each vehicle when the fraction of time a vehicle operates without a passenger on board is minimized.

Extended Data Fig. 2 reports the yearly analysis of minimum fleet requirements, along with the corresponding daily number of trips. Whereas the number of daily trips clearly displays an increasing weekly pattern, the number of required vehicles remains fairly constant, with a dip on Sundays. The robustness of the fleet size despite large variation in the number of daily trips shows that the minimum fleet size can tolerate handling extra trips without needing extra vehicles. The addition of such trips certainly leads to higher vehicle utilization, as we show here. Extended Data Fig. 4 reports a breakdown of the deployed vehicles into the different phases of deployment—passenger onboard, en route to next passenger, waiting for next passenger—for a better understanding of the utilization patterns.

Extended Data Fig. 5 reports vehicle-level performance using various temporal metrics. The vehicle start and end of operation time during the day in Extended Data Fig. 5a shows that on most days, minimum fleet assignment leads to high operation times for the majority of the vehicles. The reported plots in Extended Data Fig. 5b and c on some days clearly show the existence of a small fraction of under-used vehicles operating on average for less than two hours, serving what we call 'special-purpose' trips. These trips occur mostly on the weekend and are spatiotemporally isolated, meaning that their existence requires new vehicles because the existing vehicle–trip assignment cannot be rearranged to accommodate these trips successfully.

A bin-packing model to describe fleet-size scaling. As shown in Extended Data Fig. 6, for a large number of days with daily trips ranging from 350,000 to 550,000, there is only a small variation in the minimum fleet size. This pattern seems a bit counter-intuitive at first glance, because basic logic implies that an increase in the number of trips should somehow lead to increase in fleet size. Outside this range this expected increasing pattern holds and for a smaller number of trips we have a more-or-less linear scaling (see the result of supersampling in Extended Data Fig. 6c).

To explain the saturation pattern observed in Fig. 2, we use a simple bin-packing model to show that the reason for fleet-size robustness within a certain range is related to an existing spatiotemporal capacity to accommodate more trips in the minimum fleet. Consider a set of N vehicles with a fixed spatiotemporal capacity to accommodate k trips during a day. The exact value of k depends on the average duration of a trip during a given time of the day, and it is limited by an strict upper bound equal to 24 h on the maximum vehicle operation time. We start with a configuration where we have a certain number of trips Nx (where $x \ll k$) randomly distributed in the bins with a Poisson distribution. We start to add one trip at a time and randomly sample a small subset of n vehicles as candidate set (n is a hyperparameter of the model that we assume to be either 1 or 2). Two scenarios are possible: (1) a subset of the selected vehicles still have the capacity to accommodate more trips, in which case we randomly select one of them and assign the trip to that vehicle; (2) none of the vehicles have spatiotemporal capacity to accommodate the new trip, in which case we add a new vehicle to the system to accommodate the new trip. We repeat this process and model the relationship between the number of vehicles and number of trips in this manner.

An interesting plateau-then-increase pattern emerges from this model, which implies that for some intermediate ranges, the fleet size first increases and then shows some robustness with respect to a further increase in the number of trips, consistent with the observed pattern as observed from our minimum fleet optimization approach in Fig. 2. This simple model suggests that the reason for the minimum fleet-size robustness is that the probability of finding a vehicle which can successfully accommodate that new trip is still relatively high as many cars operate with a large unused spatiotemporal capacity when the number of trips is relatively low. The range of minimum fleet-size tolerance is determined by the maximum number of trips that a certain number of vehicles can serve in theory. This maximum number depends on the spatiotemporal distribution of trips, especially the distribution of the trip durations. For instance, if the average trip duration in a day is 10–15 min, a vehicle can serve up to around 3–4 trips per hour assuming a 5-min connection time between the trips on average. In this way the upper bound would be around 100 trips for vehicles that are active for most of the day. With this assumption the maximum number of trips a minimum fleet of around 6,000 vehicles can tolerate is around 600,000 trips. Figure 2 and the results of the model in Extended Data Fig. 6d support this argument.

Although the model in this section is an oversimplification and does not consider the complex spatiotemporal constraints that determine whether a vehicle can serve a trip, it does, however, capture the saturation pattern represented in Fig. 2. Extended Data Fig. 7 supports the idea that the robustness of the fleet size is due to the existing capacity in vehicles by showing how the metrics associated with vehicle utilization show a consistent increase in vehicle utilization for days with higher numbers of trips. Days with higher numbers of trips score higher average utilization per vehicle as can be seen in both the increase in the average time a vehicle spends on the road with a passenger on board for each day (see Extended Data Fig. 7a) and also in the increase in the average time vehicles spend waiting to pick up a passenger at the pick-up point (see Extended Data Fig. 7b).

Multi-operator model. As briefly discussed in the main text, consider a situation in which there is more than one mobility operator, each having access only to a subset of trip demand data and assuming that the operators assign the vehicles in their fleet to the trip demands they have access to without sharing information

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

with the other mobility operators. The question is to what extent the fleet size is affected by the lack of information sharing between a certain number of mobility operators. This is equivalent to going from a global optimum to a local optimum, in which each vehicle receives limited information about adjacent trips and tries to maximize its utilization independent of other vehicles. This latter situation is the extreme limit at which the number of operators is very large and only a local optimum can be achieved. In the following, using a simplified model we try to address the cases for two and three mobility operators equally sharing the mobility market.

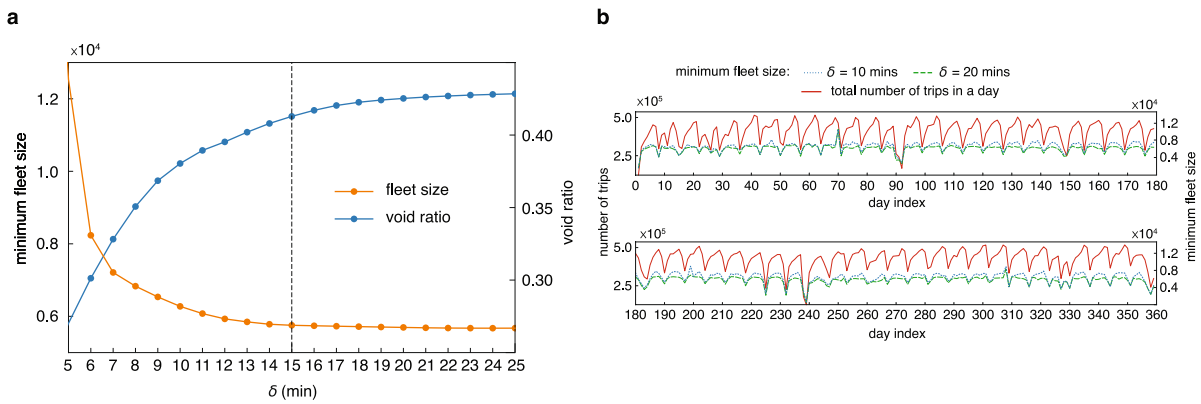
For this purpose we randomly sample the trip demand data at each given point in time and divide the trip set into multiple subsets. For each trip subset we can build a vehicle-shareability network and do the minimum fleet optimization as described in this Letter. Each optimization leads to a minimum fleet size for each mobility operator. By comparing the sum of the fleet sizes for the multiple mobility operator case with the global minimum fleet size we can find out how far away we are from the global optimum.

Extended Data Fig. 8 shows the temporal pattern of the sum of fleet sizes for a sample of 100 days from New York City taxi trip data. To obtain a good estimate for the sum of fleet sizes, we have divided the trip set in each day into two and three equally sized subsets by random subsampling. We repeated the random subsampling several times and each time we perform the vehicle-shareability network optimization to find the fleet size for each subset. The average fleet size obtained from several random subsamplings each day is then presented in Extended Data Fig. 8a and b. As shown in Extended Data Fig. 8b, the transition from a monopolistic to a oligopolistic market incurs a small drop in efficiency quantifiable at about 4%–6% for two-operator markets, and about 6%–10% for three-operator markets. A further increase in the number of operators leads to higher inefficiency in terms of fleet size as one is moving away from the global optimum achievable in the monopolistic market to an increasingly partial one. If the number of disjoint operators increases further the total size of the fleet would keep increasing owing to the lack of communication between mobility operators, even in the case when each of them try to optimize their fleet size based on the information about trip demand they receive. The fact that considering two or three operators sharing equal shares of the mobility market results only in a small drop in efficiency in terms of the fleet size shows that the minimum fleet-size optimization using the network-based approach for two or three independent operators is not far from the global optimum.

Fleet-size inflation due to rare events. The analysis of historical data has shown that our model provides a robust improvement (the reduction in fleet size) on previous models. However, an inflation of the optimal fleet size could occur as a result of rare, unusual demand patterns. For instance, if a sudden burst in the number of trips occurs around a given location but with diverging destinations, these trips will not be connected to each other in the vehicle-shareability network. Thus, for such cases to be served, the trips require separate vehicles from the existing pool on the road or even extra vehicles. These special events inflate the number of vehicles required because the nodes added to the vehicle-shareability network can have sparse or no connectivity to other nodes in the network. Although rare, such cases of trip-demand bursts can occur after events such as sports matches or concerts. However, based on our historical analysis it is evident that these outlying patterns only rarely lead to any inflation in the number of vehicles required to serve all the demand.

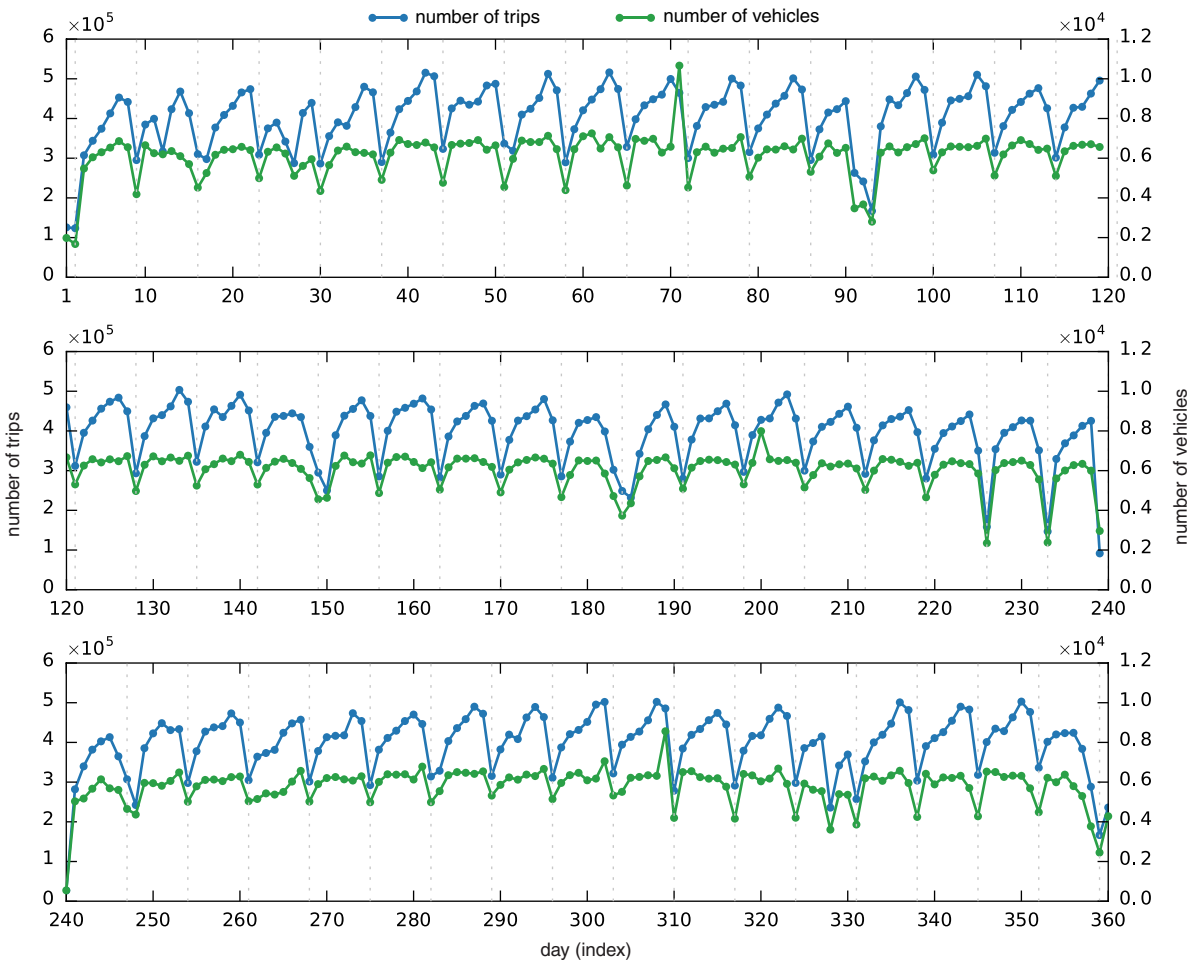
Data and code availability. All data processed during the course of this study are included in this Letter and its Supplementary Information. The code for generating the shareability networks and optimal dispatching is subject to licensing and could be made available upon request to the authors. New York City taxi data used in the study can be downloaded at http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



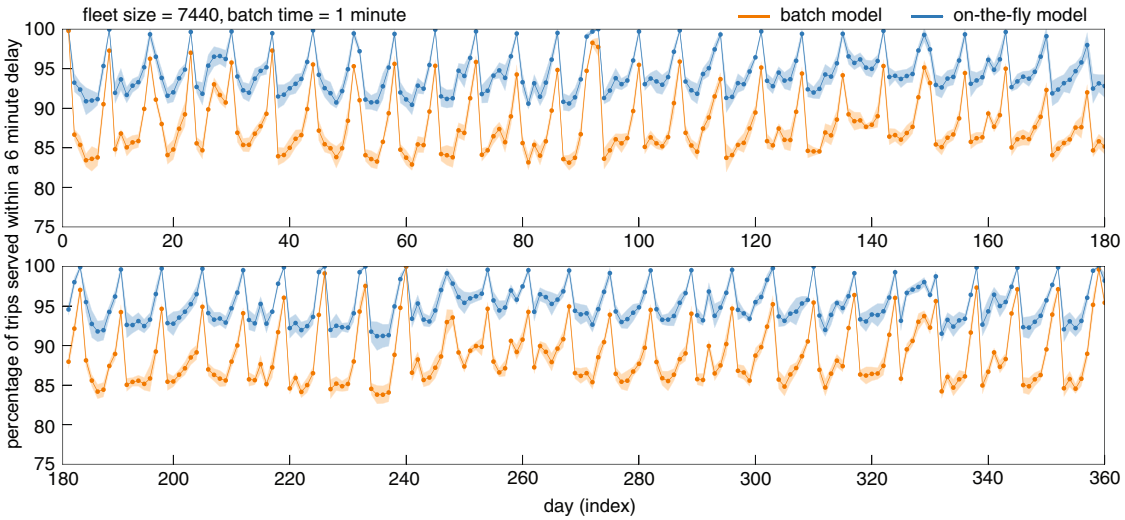
Extended Data Fig. 1 | The effect of the upper bound δ on trip connection time. a, Plot showing the minimum fleet size and the vehicle void ratio as a function of the upper bound used on the trip connection time, δ. For increasing values of δ, the minimum fleet size decreases while the void ratio increases. The results are produced by averaging over 14 days of data. **b**, Plot showing the yearly variation in the minimum fleet size for $\delta = 10$ min and 20 min. This plot reports also the number of daily trips for comparison.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



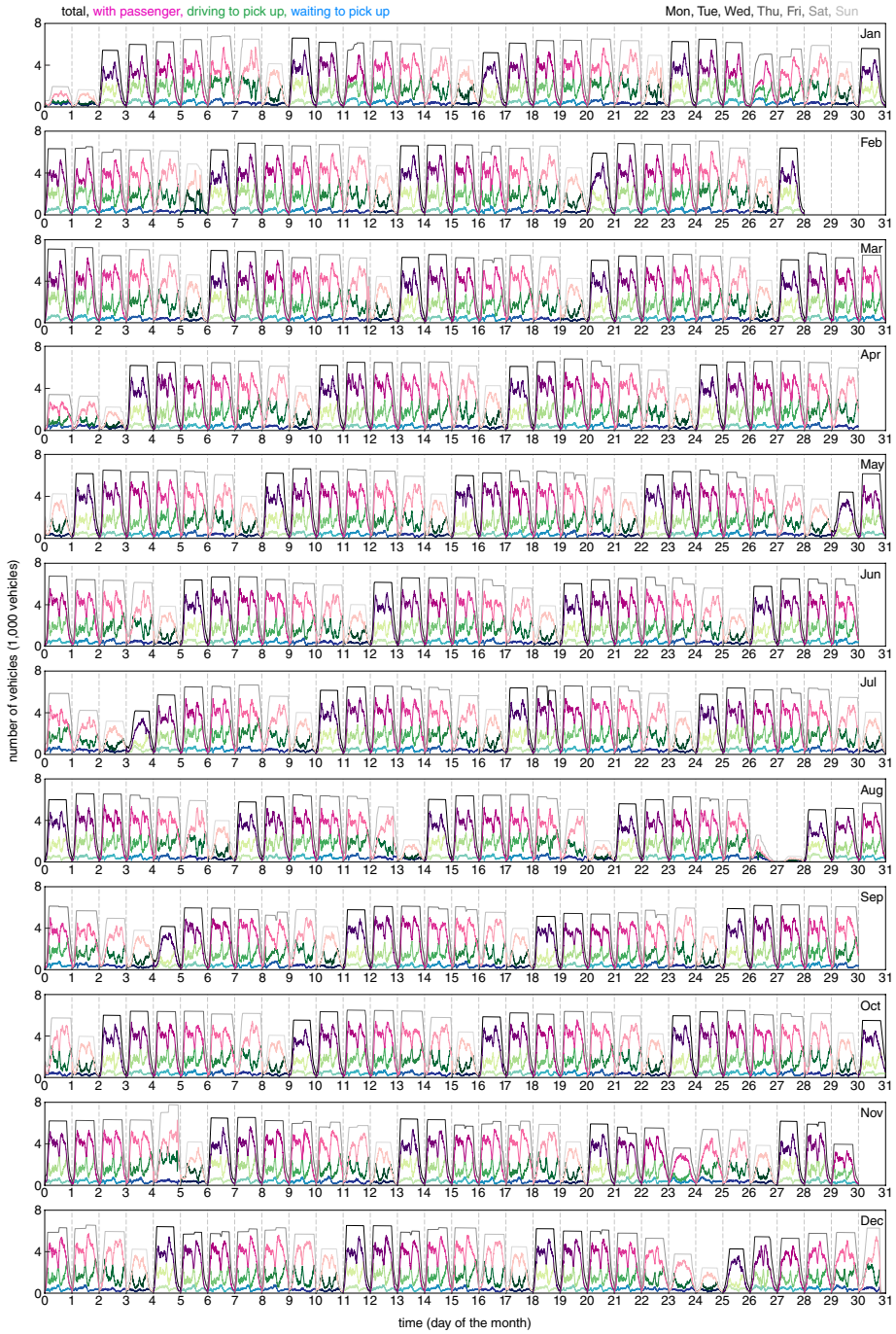
Extended Data Fig. 2 | Intra-annual patterns. For each day in year 2011, the number of daily trips (blue line) and the corresponding minimum fleet size (green line) are shown. Both quantities follow different weekly patterns: whereas the number of daily trips (blue line) consistently increases from Monday to Friday and drops on Sunday, the minimum number of vehicles needed to serve those trips (green line) remains substantially constant at a value of around 6,200, with a considerable drop on Sundays, when this value is reduced to about 4,000.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



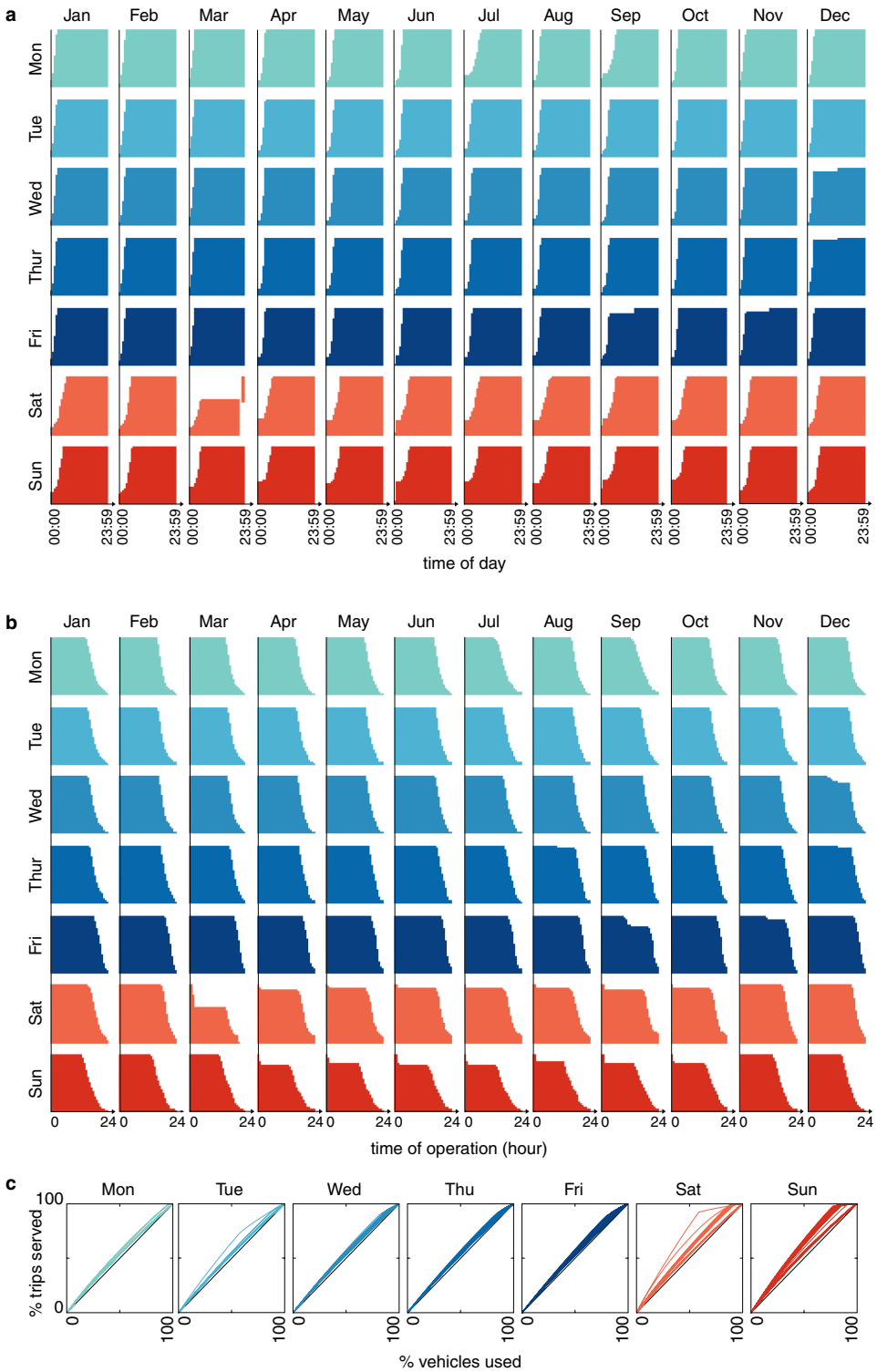
Extended Data Fig. 3 | Intra-annual comparison between batch and on-the-fly models. Plots showing the percentage of trips served within the next 6 min from the time the trip requests are received. In the batch model, advance knowledge of the trip information is restricted to only the next minute (batch time). The trips in each batch are assigned to the available vehicles using the online version of the network approach from the minimum fleet of size 7,440. This approach scores a consistently higher percentage (90%) than does the on-the-fly model. In the on-the-fly model, trips are assigned to the closest available vehicle. To achieve the same level of service using the on-the-fly model, the fleet size must increase by more than 30% (see Fig. 4b). The shaded region represents the 6σ variations when the vehicle warm-up phase is reinitialized 50 times, where $\sigma = \max(\%) - \min(\%)$ is the difference between the percentage of served trips achieved for the runs that score maximum and minimum values for each day.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 4 | Detailed temporal patterns of the minimum fleet. At each time during a day, each active vehicle in the minimum fleet set operates in one of three possible modes: (1) empty of passengers and on the way to pick up a passenger, (2) empty of passengers and waiting at a passenger's pick-up location to pick up, (3) serving, with a passenger on board. The number of vehicles operating in each of these modes computed for each minute during the day follows regular daily and weekly patterns, as shown by the three coloured curves for all months in the year. The total fleet size active on the road (black-to-grey curves) demonstrates robustness, because most of the vehicles in the minimum fleet are active at all times during the day (see also Extended Data Fig. 5a and b). Different panels correspond to different months, and the colour intensity is used to differentiate different days of the week.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

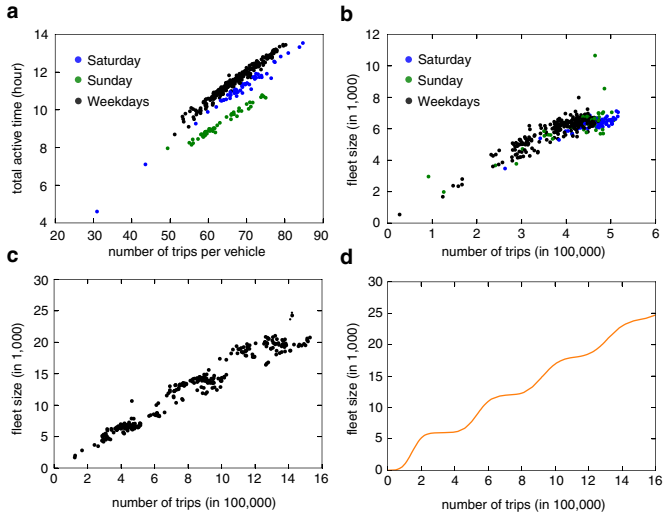


Extended Data Fig. 5 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

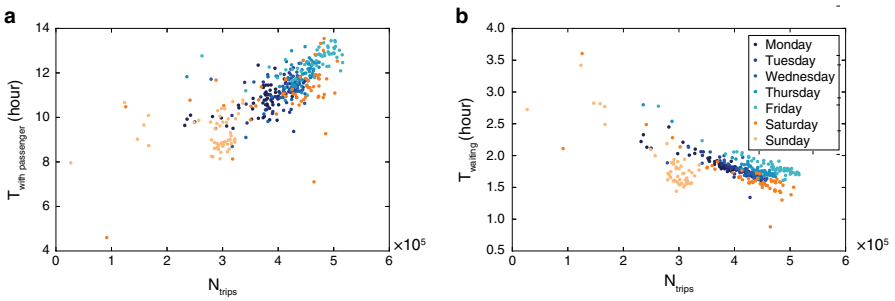
Extended Data Fig. 5 | Vehicle-level performance in the minimum fleet assignment. a, Stacked horizontal bar plots showing the start and end of the operation time (left and right ends of the stacked bars) for each vehicle in the minimum fleet assignment for various days of the week. The day in each panel represents results for days in the second week of each month. Vehicle active times are represented by a very thin coloured bar. The vehicles are sorted by the start of operation time, and stacking them horizontally creates each plot for each day. In all days (except for the outlier day of the second Saturday in March 2011), the patterns show high efficiency, with the majority of vehicles starting early in the day and operating until the end of the day. **b**, Stacked horizontal bar plots, this time representing the total operation time of the vehicle (the length of the bar). The bars are sorted based on the vehicle's total operation time, the lowest bar corresponding to the vehicle with the longest operation time. A distinct pattern emerges on most of the weekends and on some days during the week. A substantial percentage of vehicles on most weekends operate for a short time to serve a small subset of trips, which we refer to as special-demand trips. We believe that the existence of these trips requires additional vehicles because of the way their pick-up and drop-off times and locations are distributed spatiotemporally. **c**, The $q-q$ plot (where the quantile is the fraction of points below the given value) showing the percentage of trips served (vertical axis), using the vehicle-shareability minimum fleet optimization, with the percentage of vehicles represented on the horizontal axis. Vehicles are sorted on the basis of their total operation time, that is, the vehicles with longer operation times appear to the left of those with shorter operation times on the horizontal axis of these plots. Each panel corresponds to a day of the week and the curves in each panel represent all such days in the entire year (for example, all Mondays). On most weekends and consistent with the patterns observed in **b**, a large percentage of vehicles (between 5% and 10%) serve only a very small percentage of trips ($<1\%$). This can be observed from the cusps appearing near the top of some of the curves.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



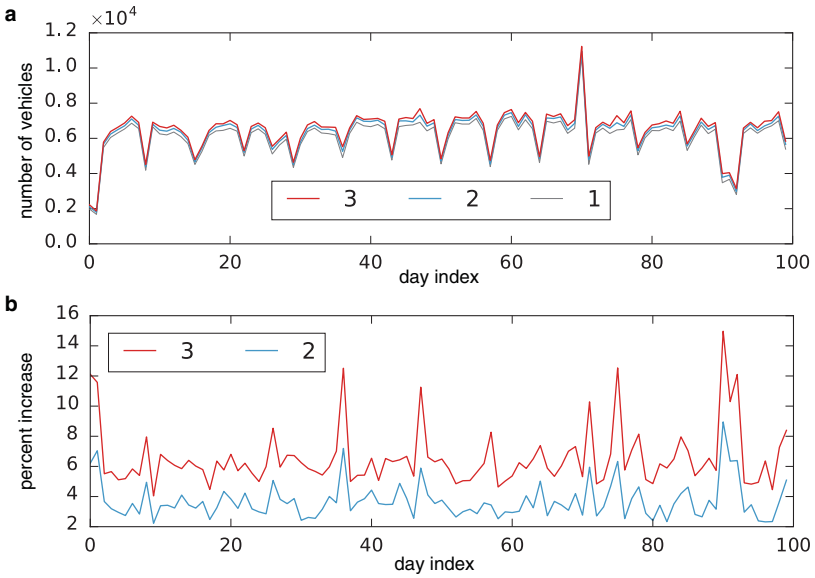
Extended Data Fig. 6 | Modelling minimum fleet size scaling with the number of trips. a, Scatter plot showing the average operation time of vehicles with optimal dispatching for different days versus the average number of trips per vehicle for each day. The former quantity scales linearly with the average number of trips per vehicle. This holds despite the fact that the fleet size manifests a saturation pattern as the number of trips grow. **b**, The coefficient of proportionality between the two quantities in **a** is different and separates out the weekends. The coefficient is slightly lower for Saturdays (blue) and much lower on Sundays (green) compared to that of weekdays. **c**, Plot showing the interplay between the minimum fleet size and the number of trips for each simulated day to manifest how the fleet size changes as the number of trips greatly increases. The supersampling is done by combining the demand for similar days in two and three successive weeks. The number of vehicles shows linear growth with a ripple-like pattern of saturation and increase. **d**, Plot showing the interplay between the fleet size and the number of trips, as simulated using a simple bin-packing model. The oversimplified model described in Methods can still capture the ripple-like saturation/increase pattern.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | Average vehicle utilization in the minimum fleet assignment versus the number of daily trips. a, Scatter plot showing the average total time with a passenger on board per vehicle in a day versus the total number of daily trips for that day. Each point in the scatter plot represents a day. The average total time with a passenger on board, which is a measure of vehicle utilization in the minimum fleet assignment, shows an overall increasing pattern with the increase in the number of daily trips that is consistent with the fact that the minimum fleet size shows robustness. **b**, Scatter plot showing the average total waiting time to pick up passengers per vehicle versus the number of daily trips for each day. The average total waiting time decreases as the number of daily trips increases, which again can be interpreted as an increase in the utilization of vehicles. The observed patterns justify the unused capacity assumption used to develop the bin-packing model (see Methods and Extended Data Fig. 6).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 8 | Efficiency comparison between single and multiple mobility operators. The optimal fleet size in the single-operator and the multi-operator mobility service in each day for the first 100 days (1 January corresponds to day index 1) in the year 2011. In the case of multiple operators, trips are randomly assigned to one of the operators in equal proportions, and network-based optimization is performed by each operator independently. The number of vehicles needed by each operator are then summed and the number for each operator is shown in **a**. **b**, Fleet-size percentage increase plot showing how the transition from a monopolistic to a oligopolistic market incurs a drop in efficiency of 4%–6% for a two-operator market, and of about 6%–10% for a three-operator market. The further increase in the number of operators leads to higher inefficiency in terms of fleet size as it moves away from the global optimum achievable in the monopolistic market to an increasingly fragmented market.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Real-time computation run times in milliseconds

x	Δt (minutes)	$t_g(ms)$ average	$t_g(ms)$ max	$t_{tva}(ms)$ average	$t_{tva}(ms)$ max
1.2	2	12	27	2	4
	3	14	29	3	8
	4	15	30	4	16
	5	17	37	6	29
	6	20	49	8	46
2.0	2	26	42	5	7
	3	28	46	7	14
	4	33	51	12	28
	5	41	68	20	55
	6	50	93	32	100

Considering the day with the highest number of trips in the year, which is day 43 for the year 2011 with around 505,000 trips, we compute the breakdown of the run times for building a bipartite trip–vehicle graph, t_g, and finding the optimum trip-to-vehicle assignment, t_{tva}, by receiving the trip requests in the next minute, on the basis of the proposed online network-based batching model. The total run time $t_g + t_{tva}$ per batch remains under 100 ms for $x = 1.2$ and under 200 ms for $x = 2.0$. This shows the practicality of the proposed method from the computational point of view. We have also varied the maximum delay, Δt between 2 min and 6 min. The average is computed for all the minutes in the day, while the maximum times correspond to the batch computation with the maximum run time. The results are based on ten separate runs for the entire day, each time reinitializing the fleet deployment warm-up phase, as described in Methods. The experiments were performed on a Linux workstation equipped with an Intel Core i7-3930K central processing unit (CPU) running at 3.20 GHz and 32 GB of random access memory (RAM). For maximum fairness, the running times are based on the actual times spent running the program, not on the CPU clocks assigned to the process.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0098-y

Self-reporting and self-regulating liquid crystals

Young-Ki Kim[1], Xiaoguang Wang[1,2], Pranati Mondkar[1], Emre Bukusoglu[1,3] & Nicholas L. Abbott[1]*

Liquid crystals (LCs) are anisotropic fluids that combine the long-range order of crystals with the mobility of liquids[1,2]. This combination of properties has been widely used to create reconfigurable materials that optically report information about their environment, such as changes in electric fields (smart-phone displays)[3], temperature (thermometers)[4] or mechanical shear[5], and the arrival of chemical and biological stimuli (sensors)[6,7]. An unmet need exists, however, for responsive materials that not only report their environment but also transform it through self-regulated chemical interactions. Here we show that a range of stimuli can trigger pulsatile (transient) or continuous release of microcargo (aqueous microdroplets or solid microparticles and their chemical contents) that is trapped initially within LCs. The resulting LC materials self-report and self-regulate their chemical response to targeted physical, chemical and biological events in ways that can be preprogrammed through an interplay of elastic, electrical double-layer, buoyant and shear forces in diverse geometries (such as wells, films and emulsion droplets). These LC materials can carry out complex functions that go beyond the capabilities of conventional materials used for controlled microcargo release, such as optically reporting a stimulus (for example, mechanical shear stresses generated by motile bacteria) and then responding in a self-regulated manner via a feedback loop (for example, to release the minimum amount of biocidal agent required to cause bacterial cell death).

We use nematic phases as representative LCs. Nematic LC phases comprise molecules that assume a preferred orientation, called the director $\boldsymbol{n}$[1]. This leads to anisotropic optical properties (birefringence) and elasticity, the latter allowing energy to be stored in strained states of LCs at rest (for example, via bending, splaying or twisting)[1,2]. When a dispersed microphase of an immiscible liquid, solid or gas (guest) is introduced into an LC host, the ordering of LCs around the guest microphase is determined by a competition between the elastic energy (KR) arising from the strain of the LC and an orientation-dependent interfacial energy at the LC–guest interface (WR^2), where K is the Frank elastic constant of the LC, W is the surface-anchoring energy density and R is the radius of the guest[2,8,9]. If $WR^2 > KR$ (that is, $R > K/W$), the guest will elastically strain the LC and thus can generate a strong repulsive force (elastic repulsion, F_e) between the guest and the boundaries of the nematic LC (Fig. 1a)[10–12]. The strain and associated topological defects will also mediate interactions between guest microphases, which can prevent contact or coalescence of liquid microphases[10–14]. For typical nematic LCs formed from small organic molecules (thermotropic LCs), $K \approx 10^{-12}$ N and $W \approx 10^{-6}$ J m^{-2} (ref. [8]); thus, $K/W \approx 1\,\mu$m and the associated elastic energy is about $2{,}400 k_B T$ (Methods), where k_B is the Boltzmann constant and T is the temperature. In the absence of external forces, therefore, we predict that dispersed microphases with $R \geq K/W$ would be sequestered by elastic forces within a bulk nematic phase and prevented from escaping into the surrounding environment.

To demonstrate the elastic trapping of guest microphases in LCs, we dispersed aqueous microdroplets ($0.5\,\mu\text{m} \leq R \leq 3\,\mu\text{m}$) containing a model chemical solute (a red dye for visualization) and the surfactant sodium dodecyl sulphate (SDS) in nematic 4′-pentyl-4-biphenylcarbonitrile (5CB; Fig. 1a), and then filled a mini-well (3.5 mm in depth) with the dispersion (Fig. 1b). The SDS adsorbed at the interface between the aqueous microdroplets and the LC and aligned $\boldsymbol{n}$ perpendicular to the local droplet interface; this is called homeotropic anchoring[7]. Accordingly, each aqueous microdroplet was surrounded by a region of strained LCs that included a point topological defect (called a hyperbolic hedgehog[10–14]; Fig. 1a). At the interface between the LC and the overlying bulk aqueous phase, the LC adopted a parallel orientation (planar anchoring), leading to a bright optical appearance of the system (left inset in Fig. 1e). As shown in Fig. 1c and Extended Data Fig. 1, the aqueous environment contacting the LC remained free of red tracer after four days at room temperature. This result is consistent with the effects of the elastic repulsion[10–12] of the guest microdroplets away from the interface between the LC and the overlying aqueous phase by a force $F_{e1} = -A^2 B \pi K [R/(h_1 + R)]^4$, where A is a numerical factor ($A = 0$ for $R < K/W$ or for an isotropic phase[10,11]), B is an anchoring-dependent constant ($B = 3/4$ and $B = 1/2$ for the parallel and perpendicular orientation, respectively, of $\boldsymbol{n}$ at the interface between the LCs and the bulk aqueous phase) and h_1 is the distance between the microdroplet surface and the LC interface to the bulk aqueous phase (Fig. 1a). The trapping of microdroplets within LCs was observed to occur regardless of the relative density of the microdroplets (ρ_{aq}) and the LCs (ρ_{LC}) because F_{e1} is much larger than buoyant forces (F_b) at room temperature; for $R = 3\,\mu$m and 5CB, $|F_{e1}/F_b| > 500$ at $h_1/R < 1$ (Methods).

Thermal and chemical stimuli can trigger optical responses in LCs via changes in LC ordering[2,4,6–8,15]. For example, 5CB undergoes a first-order transition from a nematic (N) to an isotropic (I) phase at temperature $T_{NI}^{5CB} = 35$ °C. Alternatively, isothermal phase transitions can be induced by absorption of solutes[16]. We found that heating of 5CB containing microdroplets to $T_h > 35$ °C from below (through contact with a warm body) led not only to an N-to-I phase transition accompanied by an optical response of the LCs (bright to dark; see insets and dashed line in Fig. 1e), but also to the ejection of microdroplets and red tracer into the overlying aqueous environment (Fig. 1d and solid line in Fig. 1e). The release occurred irrespective of the relative magnitudes of ρ_{aq} and ρ_{LC}, including in conditions under which the microdroplets sediment downwards and away from the interface to the overlying aqueous environment ($\rho_{aq} > \rho_{LC}$). Surprisingly, the release was transient, coinciding with the time period of the phase transition (Fig. 1e). After the phase transition, we did not measure any additional release for a period of 24 h (Extended Data Fig. 1). A second pulse of solute was released when the system was cooled back to $T_c = 25$ °C to reform the birefringent N phase (Fig. 1f). We subsequently repeated the heating and cooling cycles and observed that, with each optical response, a well defined pulse of microdroplets was ejected into the overlying aqueous phase (Fig. 1g). After 20 cycles, the mass of solute dispensed into the aqueous environment was linearly proportional to the initial concentration of aqueous microdroplets, C_{aq}, in the LCs (Fig. 1h, i), revealing a high level of control. Thermal release was also conveniently initiated by ohmic heating of a thin, electrically resistive film supporting the LCs (Extended Data Fig. 1).

We determined that the pulsatile ejection of microdroplets accompanied the upward motion of the N–I interface towards the overlying aqueous environment (Extended Data Fig. 1). Microscopic observations revealed that repulsive elastic interactions between the aqueous

[1]Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI, USA. [2]Present address: John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [3]Present address: Chemical Engineering Department, Middle East Technical University, Ankara, Turkey. *e-mail: nlabbott@wisc.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

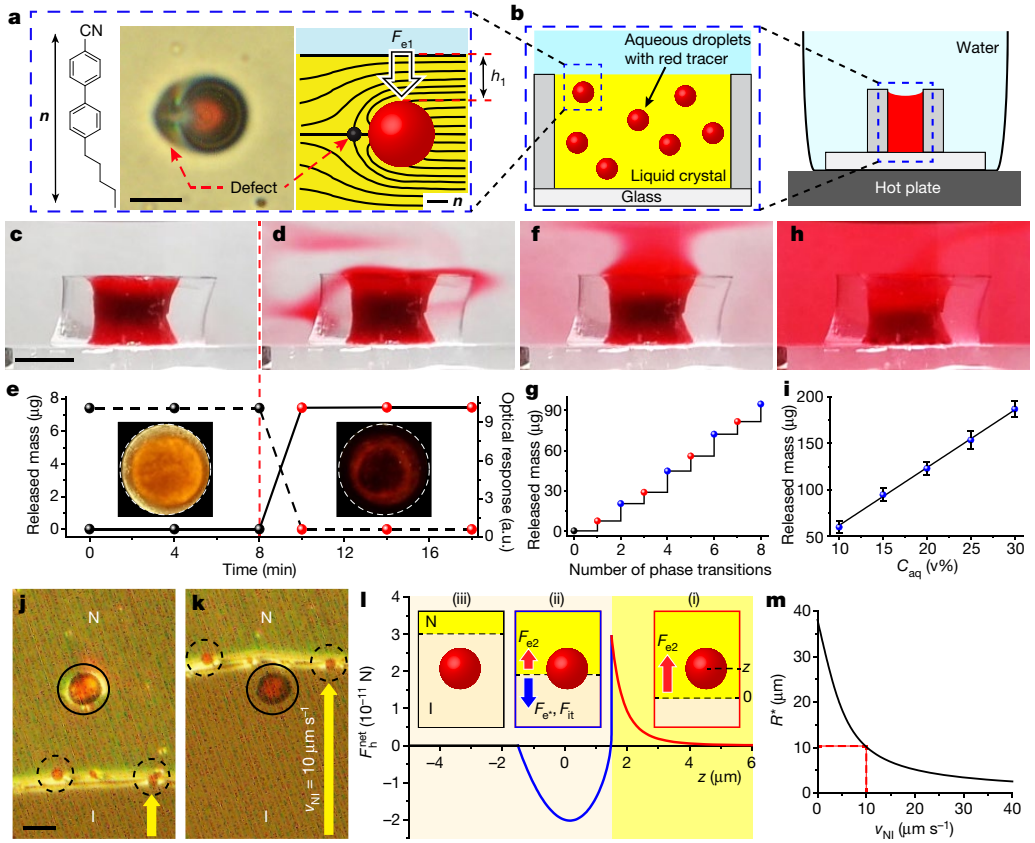


Fig. 1 | Pulsatile ejection of microdroplets from LCs and accompanying LC optical responses, as triggered by N–I phase transitions. a, Molecular structure of 5CB and optical micrograph (bright field) of an aqueous microdroplet (containing 9 mM of SDS and red tracer) in nematic 5CB with a reconstructed LC director profile. F_{e1} is the repulsive elastic force acting between the aqueous microdroplet and the interface between the LC and the overlying aqueous phase. Scale bar, 10 μm. **b**, Illustration of a dispersion of aqueous microdroplets in an LC that is hosted in a mini-well and submerged under a bulk aqueous phase that is initially free of tracer. **c**, **d**, **f**, **h**, Sequential photographs of the LC-filled mini-well containing the initial N phase (**c**) and after N-to-I (**d**), N-to-I-to-N (**f**) and 10 phase transitions (**h**); see Supplementary Video 1. The phase transitions were triggered by heating and cooling between 25 °C and 50 °C. Scale bar, 3 mm. **e**, Mass of tracer released (solid line) from LC as a function of time before (black circles) and after N-to-I phase transition (red circles); the accompanying optical response of the LC is shown by the dashed line. The insets show micrographs of mini-wells (of diameter 3 mm) between crossed polars (top view). **g**, Mass of tracer released as a function of the number of phase transitions. Red and blue circles indicate mass released after N-to-I and I-to-N phase transitions, respectively. a.u., arbitrary units. **i**, Mass of tracer released after 20 phase transitions as a function of the microdroplet concentration (C_{aq}) initially in the LC. The data points show mean values and the error bars represent 1 s.d. ($n > 3$). **j**, **k**, Sequential micrographs showing the transport of small (dashed circles; $R < R^*$), but not of large (solid circles; $R > R^*$), microdroplets by a moving N–I interface (yellow arrows; $v_{NI} = 10\,\mu m\,s^{-1}$) upon heating. Scale bar, 20 μm. **l**, Calculated net force acting on a microdroplet ($R = 1.5\,\mu m$) near an N–I interface. The insets show microdroplets for $z \geq R$ (red line), $-R < z < R$ (blue line), and $z \leq -R$ (black line). Red and blue arrows indicate the forces that favour and oppose the ejection of microdroplets, respectively. F_{e2} is the repulsive elastic force acting between the aqueous microdroplet dispersed in the N phase and the N–I interface, which is modified by the elastic force F_{e*} and interfacial tension force F_{it} when the aqueous microdroplet penetrates that N–I interface (Methods). **m**, Calculated dependence of R^* on v_{NI}. The red dashed line indicates R^* at $v_{NI} = 10\,\mu m\,s^{-1}$, which coincides with the experimental conditions.

microdroplets and the moving N–I interface pushed the microdroplets ahead of the interface in a manner that depended on the size of the microdroplets (Fig. 1j, k). Past studies[17] have observed N–I interfaces to transport adsorbed colloids, but they have not reported the elastic levitation of colloids ahead of an N–I interface, as is required for triggering the release of the microdroplet shown in Fig. 1. The elastic force—evaluated as $F_{e2} = +A^2B\pi K(R/z)^4$, where $B = 3/4$ and z is the vertical position of the centre of the microdroplet relative to an N–I interface ($z = 0$)—elastically levitates the microdroplets to a height defined by the net force $F_h^{net}(z) = 0$ above the N–I interface (see Fig. 1l and Methods)[11]. In the absence of F_{e2}, the microdroplets will not be transported by the rising N–I interface. Microdroplets pushed by F_{e2} ahead of the moving N–I interface also experience a downward Stokes drag force $F_S = -6\pi\eta_{LC}v_{aq}R$, where η_{LC} is the dynamic viscosity of the LC and v_{aq} is the velocity of the microdroplet[18] (which is equal to the velocity v_{NI} of the moving N–I interface for microdroplets transported by the N–I interface). If F_S exceeds the maximum value of $F_h^{net}(z)$ (Fig. 1l), microdroplets break through the moving N–I interface. For $(z–R)/R \ll 1$, F_{e2} is independent of R, whereas F_S scales linearly with R; thus, for each value of v_{NI}, our model defines a critical microdroplet radius R^* above which microdroplets are not transported by the moving N–I interface (Fig. 1m and Extended Data Fig. 2). Our model predicts $R^* = 10.2\,\mu m$ for $v_{NI} = 10\,\mu m\,s^{-1}$ (Fig. 1m), in good agreement with our experiments (Fig. 1j, k; $R^* = 10 \pm 1\,\mu m$). Our model also predicts that microdroplets with $R = 1.5\,\mu m$ pushed by an N–I interface with $v_{NI} = 10\,\mu m\,s^{-1}$ will be able to reach within 60 nm of the interface of an overlying aqueous phase. At this separation, attractive interfacial forces, such as van der Waals and electrical double-layer forces (see below), mediate the fusion of the microdroplets with the overlying aqueous phase (Extended Data Fig. 3).

Modified versions of the above-described elastostatic model can account for the ejection of microdroplets by the upward motion of the N–I interface during cooling (Methods and Extended Data Fig. 4). In addition, the model predicts that tuning of the elastic properties of the LC, which can be achieved using light[19], temperature[19,20] and chemical additives[6,7,16], can trigger optical responses in LCs along with a continuous release of microdroplets; these predictions were verified by experiments (see Extended Data Figs. 5–7 and Methods).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

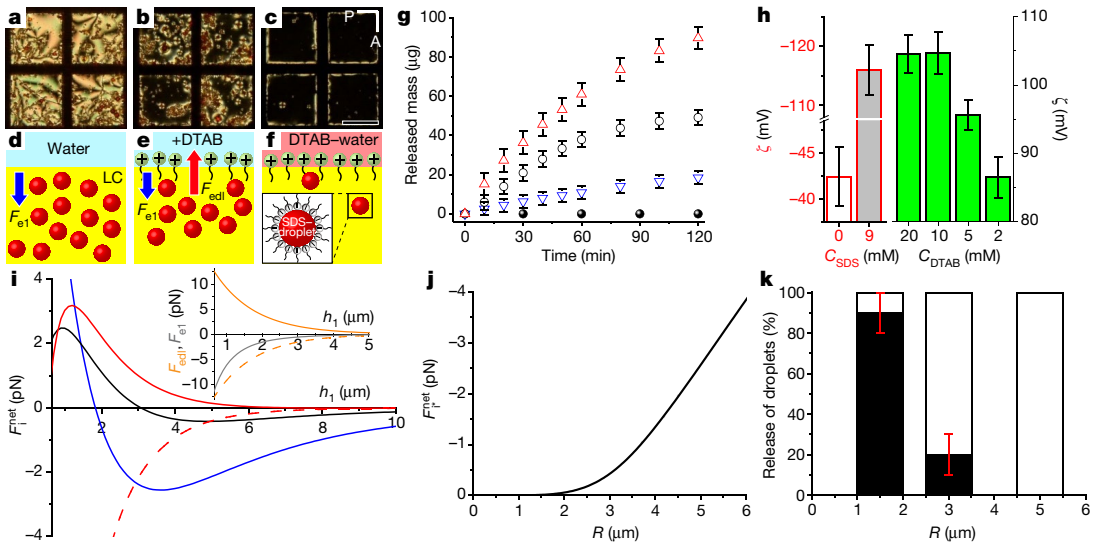


Fig. 2 | Isothermal triggering of release of microdroplets dispersed in LCs by interfacial charge interactions. a–f, Sequential micrographs (crossed polars, top view, **a–c**) and corresponding illustrations (side view, **d–f**) of an LC film (thickness 40 μm, hosted in a square grid) containing microdroplets (concentration $C_{SDS} = 9$ mM and red tracer) before (**a**, **d**) and after addition of the cationic amphiphile DTAB (10 mM) to the overlying aqueous phase, at 0 (**b**, **e**) and 60 min (**c**, **f**). P and A indicate the orientations of the polarizer and analyser, respectively. Red and blue arrows indicate the forces that favour (electrical double-layer force, F_{edl}) and oppose (elastic repulsion, F_{e1}) the ejection of microdroplets, respectively. Green (in **e**, **f**) and white (inset in **f**) circles with tails represent cationic (DTAB) and anionic (SDS) surfactants, respectively. Scale bar, 200 μm. **g**, Mass of tracer released from the LC following the addition of DTAB at concentrations $C_{DTAB} = 2$ mM (upside-down triangles), 5 mM (empty circles), 10 mM (triangles) or SDS at $C_{SDS} = 5$ mM (full circles) to the overlying aqueous phase. The symbols show mean values and the error bars are 1 s.d. ($n > 5$). **h**, Zeta potentials (ζ) of LC–aqueous interfaces without amphiphiles (white bar) and with SDS (grey bar) or DTAB (green bars). Data show mean values and the error bars are 1 s.d. ($n > 3$). **i**, Calculated net force ($F_i^{net} = F_{edl} + F_{e1}$) acting on a microdroplet ($C_{SDS} = 9$ mM) with $R = 1.5$ μm (red), 3 μm (black) and 5 μm (blue) after the addition of DTAB (10 mM; solid lines) or SDS (5 mM, $R = 1.5$ μm; red dashed line) to the overlying aqueous phase, plotted as a function of h_1. The inset shows the corresponding elastic (F_{e1}; grey) and electrical double-layer (F_{edl}; orange dashed and solid lines are for $C_{SDS} = 5$ mM and $C_{DTAB} = 10$ mM in the overlying aqueous phase, respectively) forces for $R = 1.5$ μm. **j**, Magnitude (F_{i*}^{net}) of the kinetic barrier in **i** as a function of R. **k**, Measured fraction of microdroplets released with $R = 1$–2 μm, 2.5–3.5 μm and 4.5–5.5 μm. Data are mean values and error bars are 1 s.d. ($n > 10$).

The orientations of LCs respond to interfacial interactions, which can be modulated by the adsorption of specific synthetic and biological molecules, as well as chemical transformations (for example, enzymatic or photochemical) of adsorbed species[6,7]. We found that specific interfacial events can introduce interactions (for example, interfacial charge interactions) that override the elastic trapping of microdroplets. For example, at room temperature, microdroplets containing anionic amphiphile (SDS) and solutes were elastically trapped within supported films of 5CB (thickness, 40 μm) immersed under an aqueous phase (Fig. 2a, d). The LC films have bright optical appearances due to the planar anchoring of the LCs at the interface to the bulk aqueous phase. However, addition of a cationic amphiphile (dodecyltrimethylammonium bromide; DTAB) to the bulk aqueous environment triggered an LC anchoring transition to a homeotropic orientation. The anchoring transition caused an optical response from the LC (bright to dark; see Fig. 2a, b) and the release of microdroplets into the overlying aqueous phase (Fig. 2a–g). By contrast, addition of anionic amphiphiles to the overlying aqueous phase did not trigger release of microdroplets (Fig. 2g). The rate of microdroplet release was correlated closely with the zeta potential, ζ, of the interface between the LC and aqueous phase, which is controlled by addition of charged amphiphiles (Fig. 2g, h). These observations indicate that release of microdroplets is triggered by electrical double-layer forces[21] (F_{edl}) acting between the microdroplets and the interface between the LC and the bulk aqueous phase. These forces can overcome the elastic forces (F_{e1}) that trap the microdroplet initially. We evaluated the double-layer forces as $F_{edl} = -4\pi\varepsilon_0\varepsilon(R/\lambda)(k_BT/e_c)^2 Y_p Y_i \, e^{-h_1/\lambda}$, where λ is the Debye screening length, ε is the relative dielectric constant measured along the interface normal, ε_0 is the vacuum permittivity, e_c is the elementary charge, and Y_p and Y_i are the effective surface potentials of the microdroplet and the LC interface, respectively (see Methods). The calculated net force ($F_i^{net} = F_{edl} + F_{e1}$) experienced by microdroplets near the LC interface confirmed that addition of DTAB generates an attractive F_{edl} that overcomes the repulsive F_{e1} (Fig. 2i). We also found a strong correlation between ζ, F_{edl} and the microdroplet release rate (Extended Data Fig. 8). We examined the influence of microdroplet size (R) on F_i^{net} and the experimental release of microdroplets. In the limit of $h_1 \ll R$ (close approach), larger microdroplets experience a larger attractive F_i^{net} due to F_{edl} (in this limit, F_{e1} is independent of R whereas F_{edl} scales linearly with R). Importantly, however, the distance dependence of F_{e1} and F_{edl} leads to a repulsive kinetic barrier ($F_{i*}^{net} < 0$) that grows with R (Fig. 2i, j). Specifically, our model predicts that a kinetic barrier will prevent ejection of microdroplets with $R > 2.3$ μm. Our experimental observations confirmed the existence of the size-dependent kinetic barrier (Fig. 2k). The role of interfacial charge in triggering microdroplet release was also confirmed by experiments involving changes in pH and the adsorption of polyelectrolytes (Extended Data Fig. 8). In addition, we demonstrated that adsorption of anionic bacterial lipopolysaccharides (Extended Data Fig. 9) can introduce interfacial charge interactions that trigger changes in the optical appearance of LCs and the release of microdroplets containing anti-bacterial agents (for example, DTAB) from the LCs.

Past studies[1–5] have demonstrated that interfacial shear stresses can change the orientations of LCs; here we show that this leads to changes in both $F_{edl}(\varepsilon)$ and $F_{e1}(B)$ due to the orientation dependence of ε and B, respectively. The orientation of the LC director changes when interfacial shear stresses exceed the elastic torque required to rotate the LC (which is approximately equal to $K(\partial^2\Psi/\partial d^2)$, where Ψ is the angle of the LC director from the surface normal and d is distance from the LC interface[15]). For a typical nematic LC ($K \approx 10^{-12}$ N), we calculate that an interfacial shear stress of 1 N m^{-2} is sufficient to reorient the LC director within an interfacial layer with thickness of about 10^{-6}

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

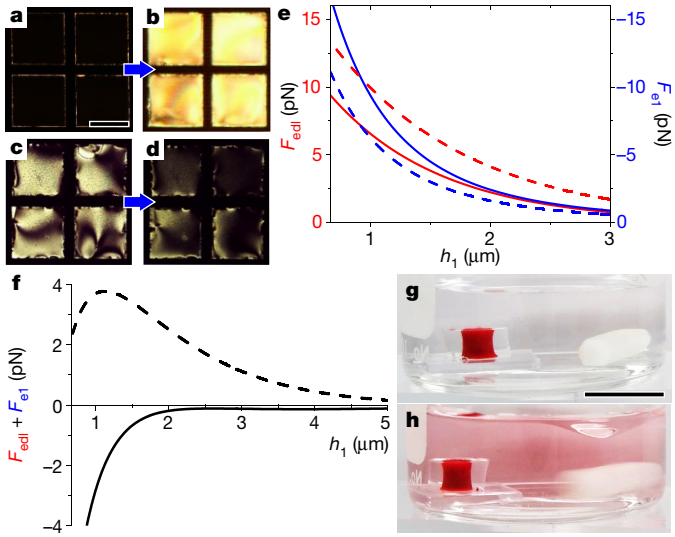


Fig. 3 | Optical response of LCs and release of microdroplets from LCs, as triggered by interfacial shear stresses. a–d, Optical response of LC films (of thickness 40 μm, hosted in a square grid supported on a treated-glass slide; crossed polars) with initially homeotropic (**a**, **b**; achieved using 2 mM SDS in the aqueous phase) or planar (**c**, **d**; surfactant-free aqueous phase) anchoring at the interface of the LCs with the overlying aqueous phase, before (**a**, **c**) and after (**b**, **d**) the introduction of interfacial shear stress via convection in the aqueous phase. Scale bar, 200 μm. **e**, **f**, Calculated F_{edl} (**e**; red lines), F_{e1} (**e**; blue lines) and net force (**f**; black lines) for an aqueous microdroplet ($R = 1.5$ μm) containing $C_{DTAB} = 2$ mM ($\zeta = 90$ mV; Fig. 2h), with planar (solid lines) or homeotropic (dashed lines) orientation of the LC at its interface (with $\zeta = -42.5$ mV) with the overlying aqueous phase. **g**, **h**, Sequential photographs showing the ejection of microdroplets ($C_{DTAB} = 2$ mM and red tracer) from the LCs, which is triggered by interfacial shear stresses, before (**g**) and after (**h**) stirring the overlying water with a magnetic bar for 30 min to generate an interfacial shear stress of about 10 N m^{-2}. Scale bar, 1 cm.

m (as seen in our experiments). In agreement with this calculation, we observed aqueous phases sheared at 10^4 s^{-1} to generate interfacial shear stresses of about 10 N m^{-2} (the viscosity of water is 10^{-3} kg m^{-1} s^{-1}) that triggered changes in the orientations and optical responses of LC films with initially homeotropic (Fig. 3a, b) or planar (Fig. 3c, d) orientations. When the initial orientation of the LC at its interface with the overlying aqueous phase was planar, we observed microdroplets containing cationic surfactant (with concentration $C_{DTAB} = 2$ mM) to be elastically trapped within the LC, consistent with our calculation that the attractive F_{edl} cannot overcome the repulsive F_{e1} (solid lines in Fig. 3e, f); we evaluated F_{edl} and F_{e1} using $\varepsilon = 13.2$, $B = 3/4$ and $\zeta = -42.5$ mV at the interface between the LC and the overlying aqueous phase (Fig. 2h). The presence of mechanical shear at the interface, however, caused the orientation of the LC to deviate from the initial planar orientation (Fig. 3c, d; see also Fig. 4), which changed the balance of orientation-dependent elastic and electrical double-layer forces. For the limiting case of a fluctuation to a perpendicular orientation ($\varepsilon = 19.7$ and $B = 1/2$), we calculated the increase in F_{edl} and decrease in F_{e1} to be sufficient to trigger ejection of microdroplets (see dashed lines in Fig. 3e, f and Methods). This prediction was confirmed by experiments showing that microdroplet release was triggered by the interfacial shear stresses (> 10 N m^{-2}) generated by shearing the overlying aqueous phase with a magnetic bar (Fig. 3g, h).

Conventional materials used for controlled release of small or large molecules do not self-report triggering events, have not been shown to respond to the same combinations of physical, chemical and biological stimuli that we describe here, and do not permit programming of the diversity of dynamic responses (for example, pulsatile or continuous release), including self-regulating responses. For example, guided by the results shown in Fig. 1, we designed a cholesteric (chiral nematic)[22] LC system that was triggered by the touch of a human finger (Fig. 4a–d; Supplementary Video 2). An N-to-I phase transition, which was designed to occur upon exposure to physiological temperature, triggered a change in the Bragg diffraction of light (as used in electronic paper[3] and LC thermometers[4]), and elastic forces F_{e2} generated ahead of the N–I interface ejected a precise dose of chemical microcargo (a dye for visualization and a cleaning agent). We note that the twisted LC within the cholesteric phase modifies the magnitude of F_e relative to an achiral nematic phase, but does not eliminate the force required to eject microcargo[23]. The programmed release of precise doses of microcargo upon thermal activation differs from the behaviour of conventional materials that release chemical agents continuously upon thermal activation until the stimulus is removed or the agent reservoir is exhausted (for example, thermally responsive hydrogels[24] and lyotropic LC matrices[25]). Current methods that release precise and repeated doses of active agent require the use of devices incorporating micrometre-scale chips, pumps, valves or flow channels[26]. The LC response shown in Fig. 4a–d can also prevent the release of excess agent (minimizing waste or toxicity) and extend the lifetime of the material so that it can withstand repeated triggering events (for example, releasing cleaning agents from a touch screen by exposure to physiological temperature). The optical response of the LC also permits self-reporting of the triggering event, which can be used to signal the amount of agent released and the remaining useful lifetime of the material.

We also used the results shown in Figs. 2 and 3 to design self-reporting and self-regulating LC materials that are triggered by mechanical shear stresses generated by the swimming motion of motile bacteria (for example, *Escherichia coli*). As noted above, interfacial shear stresses of 1 N m^{-2} can reorient an LC (see Fig. 3 and associated text) and generate LC interfacial velocities v_x of 10 μm s^{-1} (calculated from viscous stresses of about $\alpha(\partial v_x/\partial d)$ within the LC film, where α is the LC effective viscosity, $\alpha \approx 10^{-1}$ kg m^{-1} s^{-1}). We observed that the arrival of motile bacterial cells moving at 30–40 μm s^{-1} triggered shear-induced changes in the orientation of the LC, leading to optical reporting of the presence of bacteria (Fig. 4i) and triggering of the release of microcargo of antibacterial agents (Fig. 4e–h; Supplementary Video 3) via changes in F_{edl} and F_{e1} (see Fig. 3e, f). Non- ($v_x = 0$ μm s^{-1}) or weakly motile bacteria ($v_x < 10$ μm s^{-1}) generated insufficient shear stresses to trigger the LC (Extended Data Fig. 10). Whereas conventional controlled-release materials, used in areas such as healthcare, water purification and food safety, release active agents regardless of whether bacteria are present (which can lead to antibacterial resistance)[27] or not, our LC material designs are self-regulating; they do not release antibacterial agents in the absence of bacteria (Fig. 4f) and release only the minimum amount of biocidal agent required to kill the bacterial cells (Fig. 4g). Specifically, self-regulation involves a feedback loop in which the bacterial shear-triggered release of antibacterial agent causes cell death, which in turn eliminates the trigger and ends the release of antibacterial agent (Fig. 4h). In addition, the self-reporting function of the LC material signals the successful killing of the bacteria (Fig. 4i). To our knowledge, this is the first example of a material with the capability to provide a self-regulated release of chemoactive agents in response to mechanical forces generated by living cells.

Our results demonstrate that LC materials can be designed to report thermal, chemical and mechanical triggers and release microcargo in response to these triggers. Given the broad range of triggers and underlying colloidal interactions (for example, elastic, electrical double-layer and shear stresses) that are perturbed by these triggers, our results provide the basis for a highly versatile approach. For example, in addition to the results reported above, we have used LCs in free-standing forms, such as LC emulsion droplets (Extended Data Fig. 11). We have also designed LC systems that respond to combinations of stimuli (for example, thermal plus chemical), thus providing more selective responses than is possible with a single stimulus (Extended Data Fig. 11). We note that specific biological processes, including specific binding events[28] and interfacial enzymatic events[29], can lead to changes in interfacial charge, thus enabling the introduction of additional selectivity into

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

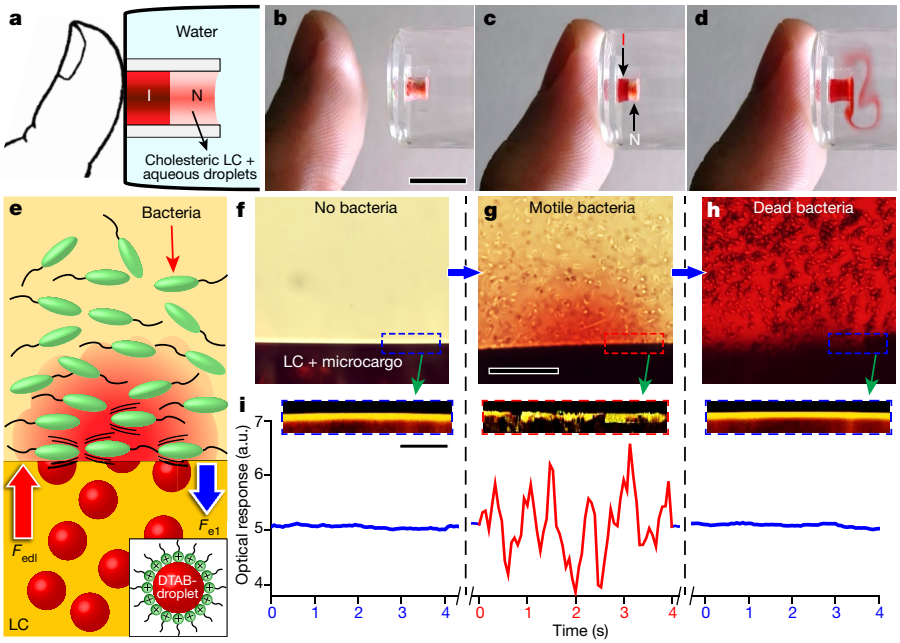


Fig. 4 | Release of microcargo from LCs by the heat of a human finger or by motile bacteria that generate interfacial shear stresses and interact with microcargo in a self-regulated manner. a–d, Schematic illustration (**a**) and sequential photographs (**b–d**) showing that the touch of a human finger can trigger a change in the Bragg diffraction of light and release of a well defined pulse of microdroplets ($C_{SDS} = 2$ mM and red tracer) from a cholesteric LC into the water (see Supplementary Video 2). Scale bar, 1 cm. **e–h**, Illustration (side view, **e**) and sequential photographs (side view, **f–h**) showing interfacial shear stresses generated by motile bacteria (10^7–10^8 *E. coli* cells per millilitre) triggering the self-regulated release of microdroplets containing anti-bacterial agent ($C_{DTAB} = 2$ mM and silver salts) and red tracer, before the arrival of bacteria (**f**), immediately after the arrival of motile bacteria (**g**; Supplementary Video 3) and two hours after the arrival of bacteria and following cell death (**h**). Scale bar, 40 μm. **i**, Optical responses and side views (crossed polars) of the LC interface corresponding to **f–h**. Scale bar, 10 μm.

the reporting and release of microcargo. Promising future directions include the use of magnetic or electric fields to modulate the responses and actions of LC systems, the use of other LC phases, such as smectic and edible lyotropic chromonic LCs[30] and the triggered release of solid microcargo from LCs (Extended Data Fig. 12).

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0098-y.

Received: 20 April 2017; Accepted: 2 March 2018;
Published online 9 May 2018.

1. de Gennes, P. G. & Prost, J. *The Physics of Liquid Crystals* (Clarendon Press, Oxford, 1993).
2. Kleman, M. & Lavrentovich, O. D. *Soft Matter Physics: An Introduction* (Springer, New York, 2003).
3. Yang, D.-K. & Wu, S.-T. *Fundamentals of Liquid Crystal Devices* (John Wiley & Sons, Chichester, 2006).
4. Demus, D., Goodby, J., Gray, G. W., Spiess, H.-W. & Vill, V. *Handbook of Liquid Crystals* Vol. 2 (Wiley-VCH, New York, 1998).
5. Cuennet, J. G., Vasdekis, A. E., De Sio, L. & Psaltis, D. Optofluidic modulator based on peristaltic nematogen microflows. *Nat. Photon.* **5**, 234–238 (2011).
6. Bukusoglu, E., Pantoja, M. B., Mushenheim, P. C., Wang, X. & Abbott, N. L. Design of responsive and active (soft) materials using liquid crystals. *Annu. Rev. Chem. Biomol.* **7**, 163–196 (2016).
7. Lin, I. H. et al. Endotoxin-induced structural transformations in liquid crystalline droplets. *Science* **332**, 1297–1300 (2011).
8. Kim, Y.-K., Shiyanovskii, S. V. & Lavrentovich, O. D. Morphogenesis of defects and tactoids during isotropic nematic phase transition in self assembled lyotropic chromonic liquid crystals. *J. Phys. Condens. Matter* **25**, 404202 (2013).
9. Wood, T. A., Lintuvuori, J. S., Schofield, A. B., Marenduzzo, D. & Poon, W. C. K. A self-quenched defect glass in a colloid-nematic liquid crystal composite. *Science* **334**, 79–83 (2011).
10. Chernyshuk, S. B. & Lev, B. I. Theory of elastic interaction of colloidal particles in nematic liquid crystals near one wall and in the nematic cell. *Phys. Rev. E* **84**, 011707 (2011).
11. Pishnyak, O. P., Tang, S., Kelly, J. R., Shiyanovskii, S. V. & Lavrentovich, O. D. Levitation, lift, and bidirectional motion of colloidal particles in an electrically driven nematic liquid crystal. *Phys. Rev. Lett.* **99**, 127802 (2007).
12. Poulin, P., Stark, H., Lubensky, T. C. & Weitz, D. A. Novel colloidal interactions in anisotropic fluids. *Science* **275**, 1770–1773 (1997).
13. Loudet, J.-C., Barois, P. & Poulin, P. Colloidal ordering from phase separation in a liquid-crystalline continuous phase. *Nature* **407**, 611–613 (2000).
14. Musevic, I., Skarabot, M., Tkalec, U., Ravnik, M. & Zumer, S. Two-dimensional nematic colloidal crystals self-assembled by topological defects. *Science* **313**, 954–958 (2006).
15. Kim, Y.-K., Senyuk, B. & Lavrentovich, O. D. Molecular reorientation of a nematic liquid crystal by thermal expansion. *Nat. Commun.* **3**, 1133 (2012).
16. Mukherjee, P. K. The T_{NI}–T^* puzzle of the nematic-isotropic phase transition. *J. Phys. Condens. Matter* **10**, 9191–9205 (1998).
17. West, J. L. et al. Drag on particles in a nematic suspension by a moving nematic-isotropic interface. *Phys. Rev. E* **66**, 012702 (2002).
18. Loudet, J. C., Hanusse, P. & Poulin, P. Stokes drag on a sphere in a nematic liquid crystal. *Science* **306**, 1525 (2004).
19. White, T. J. & Broer, D. J. Programmable and adaptive mechanics with liquid crystal polymer networks and elastomers. *Nat. Mater.* **14**, 1087–1098 (2015).
20. Borshch, V. et al. Nematic twist-bend phase with nanoscale modulation of molecular orientation. *Nat. Commun.* **4**, 2635 (2013).
21. Mustin, B. & Stoeber, B. Single layer deposition of polystyrene particles onto planar polydimethylsiloxane substrates. *Langmuir* **32**, 88–101 (2016).
22. Zheng, Z. G. et al. Three-dimensional control of the helical axis of a chiral nematic liquid crystal by light. *Nature* **531**, 352–356 (2016).
23. Varney, M. C. M. et al. Periodic dynamics, localization metastability, and elastic interaction of colloidal particles with confining surfaces and helicoidal structure of cholesteric liquid crystals. *Phys. Rev. E* **90**, 062502 (2014).
24. Kearney, C. J. & Mooney, D. J. Macroscale delivery systems for molecular and cellular payloads. *Nat. Mater.* **12**, 1004–1017 (2013).
25. Zabara, A. & Mezzenga, R. Controlling molecular transport and sustained drug release in lipid-based liquid crystalline mesophases. *J. Control. Release* **188**, 31–43 (2014).
26. Nguyen, N.-T., Shaegh, S. A. M., Kashaninejad, N. & Phan, D.-T. Design, fabrication and characterization of drug delivery systems based on lab-on-a-chip technology. *Adv. Drug Deliv. Rev.* **65**, 1403–1419 (2013).
27. Lehar, S. M. et al. Novel antibody-antibiotic conjugate eliminates intracellular S. aureus. *Nature* **527**, 323–328 (2015).
28. Sidiq, S., Prasad, G., Mukhopadhaya, A. & Pal, S. K. Poly(L-lysine)-coated liquid crystal droplets for cell-based sensing applications. *J. Phys. Chem. B* **121**, 4247–4256 (2017).
29. Ma, C. D. et al. Liquid crystal interfaces programmed with enzyme-responsive polymers and surfactants. *Small* **11**, 5747–5751 (2015).
30. Lydon, J. Chromonic review. *J. Mater. Chem.* **20**, 10071–10099 (2010).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Acknowledgements This work was primarily funded by the Army Research Office (through grants W911NF-15-1-0568 and W911NF-17-1-0575), the National Science Foundation (grants CBET- 1263970 and DMR-1435195) and the Wisconsin Materials Research Science and Engineering Center (grant DMR-1720415). We thank R. Trivedi and D. B. Weibel for assistance in the preparation of the bacterial dispersions.

Reviewer information *Nature* thanks T. Ware and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions The experimental strategy was proposed initially by N.L.A. and developed by all authors. Experiments were performed by Y.-K.K. with assistance from X.W., P.M. and E.B. and analysed by Y.-K.K. and N.L.A. Y.-K.K. and N.L.A. wrote the manuscript with input from all co-authors.

Competing interests The University of Wisconsin-Madison has filed a patent application (PCT/US207/037414) on the work described in this manuscript. The inventors listed on the patent application are N.L.A., Y.-K.K., X.W. and E.B.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0098-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0098-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to N.L.A.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Comparison of the magnitudes of elastic and thermal energies. The elastic interaction energy E_e between a microdroplet and the boundaries of the nematic LC (for example, the interfaces of N–overlying aqueous phase, N–glass, N–I) can be written as[10–12]:

$$E_e = A^2 B\pi K \frac{R^4}{(h+R)^3} \quad (1)$$

where A is a numerical factor ($A = 2.04$ for microdroplets of $R > K/W$ with homeotropic anchoring in an N phase and $A = 0$ for microdroplets of $R < K/W$ or in an I phase)[10], B is the anchoring-dependent constant ($B = 3/4$ and $B = 1/2$ for planar and homeotropic anchoring, respectively, at the boundaries of nematic LCs), R is the radius of the guest microdroplet, h is the distance between a microdroplet surface and the nematic boundary and $K = (K_1 + K_3)/2$ is the Frank elastic constant of the LC, where K_1 and K_3 are elastic constants for splay and bend deformations, respectively[1]. As a microdroplet approaches a nematic interface, E_e increases and exhibits a maximum at $h = 0$. Because $K \approx 10^{-12}$ N for typical thermotropic LCs, the maximum elastic interaction energy E_e^{max} for a microdroplet with $R = 1\,\mu m$ and $B = 3/4$ is 9.8×10^{-18} J. $k_B T = 4.1 \times 10^{-21}$ J at $T = 25\,°C$, and thus $E_e^{max} = 2,383 k_B T$.

Comparison of the magnitudes of elastic forces and buoyant forces. Based on equation (1), the elastic repulsive force F_e between a microdroplet and a nematic interface can be written as[10,11]

$$F_e \approx \pm A^2 B\pi K \left(\frac{R}{h+R}\right)^4 \quad (2)$$

F_e is valid for a microdroplet in an N phase and needs to be modified for a microdroplet at an N–I interface (see below). The maximum F_e is $F_e^{max} = A^2 B\pi K$ at $h = 0$ and the buoyant force F_b acting on a microdroplet in the LC is $F_b = (4/3)\pi R^3 g(\rho_{LC} - \rho_{aq})$. At 25 °C, $K_{5CB} = 7.3$ pN (ref. [31]), $K_{E7} = 14.4$ pN (ref. [32]), $\rho_{5CB} = 1.010\,g\,cm^{-3}$ (ref. [33]), and we measured $\rho_{E7} = 1.057\,g\,cm^{-3}$, $\rho_{aq} = 1.018\,g\,cm^{-3}$ with red dye and $\rho_{aq} = 1.012\,g\,cm^{-3}$ with green dye. Therefore, for a microdroplet (red dye) of $R = 3\,\mu m$ and $B = 3/4$ in 5CB, we get $F_e^{max} = 8,072 F_b$ and for a microdroplet (green dye) of $R = 4\,\mu m$ and $B = 3/4$ in E7, we have $F_e^{max} = 1,194 F_b$.

Net force $F_h^{net}(z)$ upon heating (Fig. 1). The net force $F_h^{net}(z)$ acting on a quasi-static microdroplet with respect to the distance z between the centre of the microdroplet and the N–I interface ($z = 0$) upon heating can be written as

$$F_{h,N}^{net}(z) = \left[-A^2 B\pi K \left(\frac{R}{h_1+R}\right)^4\right]_{F_{e1}} + \left[A^2 B\pi K \left(\frac{R}{z}\right)^4\right]_{F_{e2}} + \left[\frac{4}{3}\pi R^3 g(\rho_{LC} - \rho_{aq})\right]_{F_b} \quad (3)$$

at $z \geq R$ (in the N phase; see Fig. 1l (i)),

$$F_{h,NI}^{net}(z) = \left[-A^2 B\pi K \frac{1}{\left(h_1 + \frac{z+R}{2}\right)^4}\left(\frac{z+R}{2}\right)^4\right]_{F_{e1}} + [A^2 B\pi K]_{F_{e2}} + \left[\frac{4}{3}\pi R^3 g(\rho_{LC} - \rho_{aq})\right]_{F_b} + \left[-2\pi R\sigma_{NI}\left(1 - \frac{z^2}{R^2}\right)\right]_{F_{it}} + \left[-\left(1+\frac{z}{R}\right)\left\{\alpha\left[WR + \frac{(WR)^2}{K}\right] + \beta K\right\}\right]_{F_{e*}} \quad (4)$$

at $-R < z < R$ (at the N–I interface; see Fig. 1l (ii)), and

$$F_{h,I}^{net}(z) = \left[\frac{4}{3}\pi R^3 g(\rho_{LC} - \rho_{aq})\right]_{F_b} \quad (5)$$

at $z \leq -R$ (in the I phase; see Fig. 1l (iii)). In these equations, h_1 is the distance between the microdroplet surface and the interface between the LC and the overlying aqueous phase (Fig. 1a), $\alpha = 0$ for $R > K/W$, $A = \beta = 0$ for $R < K/W$, and $B = 3/4$ (parallel orientation at nematic boundaries).

When a microdroplet ($R > K/W$; homeotropic anchoring) is in the N phase ($z \geq R$), the net force $F_{h,N}^{net}(z)$ arises from F_b and F_e with[10,11] $A = 2.04$; see Fig. 1l (i). As the N–I interface ($z = 0$) approaches the microdroplet, $F_{h,N}^{net}(z)$ increases and becomes positive (upward) at $z < z^*$ (at $z = z^*$, $F_{h,N}^{net} = 0$) owing to the elastic repulsion from the N–I interface (F_{e2} in equation (3)); $z^* = 17.0\,\mu m$ for a microdroplet with $R = 1.5\,\mu m$ and $h_1 = 1\,mm$ in 5CB. This reveals that the microdroplet will be levitated at $z = z^*$ above the N–I interface. $F_{h,N}^{net}(z)$ shows a maximum at $z = R$ (red curve in Fig. 1l).

When the microdroplet penetrates the N–I interface ($-R < z < R$; Fig. 1l (ii)), two additional forces are generated from the changes in interfacial tensions[17,34] (F_{it}) and elastic strain[17,35] (F_{e*}). These forces drive the microdroplets into the I phase (Fig. 1l (ii)); see separate sections on F_{it} and F_{e*} for additional detail. Additionally, F_e needs to be modified because the part of the microdroplet that protrudes into the I phase no longer strains the LC. Because R in equation (2) is the radius of a microdroplet in the N phase, for simplicity, we used $(R + z)/2$ for heating and $(R - z)/2$ for cooling as the effective radius of the part of each microdroplet in the N phase. We also need to take into account the decrease in the topological strength m of the microdroplet[2] from 1 to θ/π, where θ ($0 \leq \theta \leq \pi$) is half of the central angle of the part of the microdroplet in the N phase; see Extended Data Fig. 4a (ii). Because the elastic interaction is proportional[1,2] to m^2, the numerical factor A in equation (4) can be described as $A = 2.04\,m = (2.04/\pi)\cos^{-1}(-z/R)$. Consequently, $F_{h,NI}^{net}(z)$ decreases rapidly at $-R < z < R$ (blue curve in Fig. 1l).

When the microdroplet is in the I phase ($z \leq -R$; Fig. 1l (iii)), the net force $F_{h,I}^{net}(z)$ is composed of only F_b; $A = 0$ and thus $F_e = 0$. Therefore, the microdroplets in 5CB ($\rho_{5CB} < \rho_{aq}$) sink whereas the microdroplets in E7 ($\rho_{E7} > \rho_{aq}$) rise.

Upon heating, the elastic repulsion from the N–I interface promotes release of the microdroplets (F_{e2} in equation (3)). Therefore, the moving N–I interface can only transport microdroplets with $R > K/W$ in the nematic phase ($F_{e2} \neq 0$). The interface passes through the microdroplets with $R < K/W$ because $F_{e2} = 0$.

Net force $F_c^{net}(z)$ upon cooling (Fig. 1). The net force $F_c^{net}(z)$ acting on a quasi-static microdroplet with respect to z can be written as:

$$F_{c,I}^{net}(z) = \left[\frac{4}{3}\pi R^3 g(\rho_{LC} - \rho_{aq})\right]_{F_b} \quad (6)$$

at $z \geq R$ (in the I phase; Extended Data Fig. 4a (i)),

$$F_{c,NI}^{net}(z) = \left[A^2 B\pi K \frac{1}{\left(h_2 + \frac{R-z}{2}\right)^4}\left(\frac{R-z}{2}\right)^4\right]_{F_{e3}} + [-A^2 B\pi K]_{F_{e2}} + \left[\frac{4}{3}\pi R^3 g(\rho_{LC} - \rho_{aq})\right]_{F_b} + \left[2\pi R\sigma_{NI}\left(1 - \frac{z^2}{R^2}\right)\right]_{F_{it}} + \left[\left(1-\frac{z}{R}\right)\left\{\alpha\left[WR + \frac{(WR)^2}{K}\right] + \beta K\right\}\right]_{F_{e*}} \quad (7)$$

at $-R < z < R$ (at the N–I interface; see Extended Data Fig. 4a (ii)), and

$$F_{c,N}^{net}(z) = \left[A^2 B\pi K \left(\frac{R}{h_2+R}\right)^4\right]_{F_{e3}} + \left[-A^2 B\pi K \left(\frac{R}{z}\right)^4\right]_{F_{e2}} + \left[\frac{4}{3}\pi R^3 g(\rho_{LC} - \rho_{aq})\right]_{F_b} \quad (8)$$

at $z \leq -R$ (in the N phase; see Extended Data Fig. 4a (iii)). In these equations, F_{e3} is the upward elastic repulsion from an LC–glass interface and h_2 is the distance between a microdroplet surface and the LC–glass interface (Extended Data Fig. 5k). For $R > K/W$, $A = (2.04/\pi)\cos^{-1}(z/R)$ in equation (7) and $A = 2.04$ in equation (8); for $R < K/W$, $A = 0$; $B = 3/4$. In the I phase ($z \geq R$; see Extended Data Fig. 4a (i)), aqueous microdroplets in 5CB sink because $\rho_{5CB} < \rho_{aq}$ (that is, $F_{c,I}^{net}(z) < 0$).

In contrast to heating, the two additional forces (F_{it} and F_{e*}) at $-R < z < R$ are directed upwards upon cooling (see Extended Data Fig. 4a (ii)); $F_{it} > 0$ and $F_{e*} > 0$. As a result, $F_{c,NI}^{net}(z)$ becomes positive and exhibits a maximum at $-R < z < R$ (Extended Data Fig. 4a). Importantly, upon cooling $F_{it} > 0$ and $F_{e*} > 0$, regardless of R. This indicates that the cooling N–I interface can transport the microdroplets with both $R > K/W$ and $R < K/W$ (Extended Data Fig. 4b), whereas upon heating the interface cannot transport the microdroplets with $R < K/W$ (because $F_e = 0$).

At $z \leq -R$ (N phase; see Extended Data Fig. 4a (iii)), the microdroplets with $R > K/W$ are sequestered in a nematic bulk whereas the microdroplets with $R < K/W$ sediment away from the N–I interface.

In Fig. 1c–i, v_{NI} upon both cooling and heating was $37 \pm 3\,\mu m\,s^{-1}$, for which our model predicts the dispensing of microdroplets with $K/W \leq R \leq 3\,\mu m$ upon heating and $0.6\,\mu m \leq R \leq 6\,\mu m$ upon cooling. This prediction is consistent with our observation that the amount of tracer released upon cooling was greater than that released upon heating (Fig. 1g).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Interfacial tension force, F_{it}. When a microdroplet is near an interface dividing two immiscible fluids (N and I phase in our case), F_{it} arises to minimize the surface tension energy[34]. Typically, microdroplets are stabilized at the interface between two immiscible fluids because the interface tensions are similar in magnitude to each other. In thermotropic LCs, however, the surface tension at the N–I interface (σ_{NI}) is much smaller than the surface tension at the interfaces between the aqueous solution and the N (σ_{aq-N}) or I (σ_{aq-I}) phases. In the case of 5CB, for example, at $T = 35\,°C$, $\sigma_{aq-N} \approx 7 \times 10^{-3}$ J m^{-2}, $\sigma_{aq-I} \approx 6 \times 10^{-3}$ J m^{-2} and $\sigma_{NI} \approx 10^{-5}$ J m^{-2} ($\sigma_{aq-N} > \sigma_{aq-I} \gg \sigma_{NI}$)[33,36]. As a result, the aqueous microdroplets at the N–I interface are expelled to the I phase[34]. For simplicity, we assume that F_{it} is active only when the microdroplet contacts the N–I interface; $F_{it} = 0$ at $|z| \geq R$.

Elastic force, F_{e*}. When a microdroplet penetrates the N–I interface ($-R < z < R$), the elastic force acting on the microdroplet is modified[17,35] by F_{e*}. Whereas F_e acts to keep microdroplets in the nematic phase, F_{e*} expels the microdroplets into the isotropic phase to minimize the elastic free energy. In the weak anchoring regime ($R < K/W$), F_{e*} originates from the anchoring of the director at the microdroplet surface and the director deformation in the bulk nematic phase, and can be written as

$$F_{e*}(R < K/W) = [WR\, f_1(z/R)]_{surface} + \left[\frac{(WR)^2}{K} f_1(z/R)\right]_{bulk} \quad (9)$$

where $f_1(z/R)$ is a dimensionless function of the penetration depth (z/R) of a microdroplet into an N phase[17,35].

In the strong anchoring regime ($R > K/W$), F_{e*} is given by

$$F_{e*}(R > K/W) = [K\, f_2(z/R)]_{bulk} \quad (10)$$

where $f_2(z/R)$ is also a dimensionless function[17,35]. Andrienko et al.[37] found that the force acting on a particle passing through an N–I interface is linearly proportional to the penetration depth z/R. In our evaluation, therefore, we simplified the dimensionless functions to $f_1(z/R) = \alpha(\mp 1 - z/R)$ and $f_2(z/R) = \beta(\mp 1 - z/R)$ where – and + denote N-to-I and I-to-N phase transitions, respectively.

Electrical double-layer interaction, F_{edl} (Figs. 2 and 3). F_{edl} can be calculated as

$$F_{edl} = -4\pi\varepsilon_0\varepsilon\frac{R}{\lambda}\left(\frac{k_BT}{e_c}\right)^2 Y_p Y_i e^{-h/\lambda} \quad (11)$$

where ε_0 is the vacuum permittivity, ε is the relative permittivity of the LC, e_c is the elementary charge, Y_p is the effective surface potential of the microdroplet, Y_i is the effective surface potential of the interface between the LCs and the overlying aqueous phase and λ is the Debye screening length[21]. λ, Y_p and Y_i can be written as

$$\lambda = \sqrt{\frac{\varepsilon_0\varepsilon k_BT}{2e_c^2 N_A I}} \qquad Y_p = \frac{8\tanh(D\zeta_p e_c/k_BT)}{1 + \sqrt{1 - \frac{2(R/\lambda)+1}{((R/\lambda)+1)^2}\tanh^2\left(\frac{D\zeta_p e_c}{k_BT}\right)}}$$
$$Y_i = 4\tanh\left(\frac{D\zeta_i e_c}{k_BT}\right) \quad (12)$$

where N_A is the Avogadro constant ($N_A = 6.02 \times 10^{23}$ mol^{-1}), I is the ionic strength of the LC, D is a numerical factor and ζ_p and ζ_i are the zeta potentials at the microdroplet surface and LC interface, respectively. In our evaluation, we used D values between 1 and 2 and $I = 1.8 \times 10^{-5}$ mol m^{-3} (corresponding to $\lambda \approx 0.66\,\mu$m at the LC interface with planar anchoring); $\lambda \approx 1.5\,\mu$m in 5CB without added electrolyte[38]. The range of D values reflects the fact that the surface potential ($\sim|D\zeta|$) is typically greater than the corresponding zeta potential[39] ($|\zeta|$). We note that the exact value of D does not change the conclusions of our model. The value of D will influence the calculated radius of the microdroplets above which the kinetic barrier (F_{i*}^{net}) will prevent release of the microdroplets. Because equation (11) is valid at $h > \lambda$ and it tends to overpredict[21] F_{edl} at $h < \lambda$, we only evaluated F_{edl} at $h \geq \lambda$.

In Figs. 2 and 3, therefore, the net force F_i^{net} can be written as

$$F_i^{net} = \left[-4\pi\varepsilon_0\varepsilon\frac{R}{\lambda}\left(\frac{k_BT}{e_c}\right)^2 Y_p Y_i e^{-h_1/\lambda}\right]_{F_{edl}} + \left[-(2.04)^2 B\pi K\left(\frac{R}{h_1+R}\right)^4\right]_{F_{e1}} \quad (13)$$

As shown in Fig. 2a–c, microdroplets containing anionic SDS ($C_{SDS} = 9$ mM) elastically sequestered in LCs can be ejected by an upward F_{edl} (> 0) induced by the addition of cationic DTAB ($C_{DTAB} = 10$ mM) into an overlying phase; for a microdroplet of $R = 1.5\,\mu$m in 5CB, $F_{edl} = 9.2$ pN and $F_{e1} = -6.2$ pN at $h_1 = 1\,\mu$m. As evidenced by the optical response shown in Fig. 2, SDS or DTAB causes an anchoring transition from a parallel to perpendicular orientation at the interface between the LCs and the overlying aqueous phase. Because the anchoring transition involves the variation of B and ε (for the parallel orientation $B = 3/4$ and $\varepsilon = (\varepsilon_\perp + \varepsilon_\parallel)/2$, and for the perpendicular orientation $B = 1/2$ and $\varepsilon = \varepsilon_\parallel$), F_{edl} increases, whereas F_{e1} decreases. For 5CB, $\varepsilon_\perp = 6.7$ and $\varepsilon_\parallel = 19.7$ at $T = 25\,°C$ and we used $\varepsilon = (\varepsilon_\perp + \varepsilon_\parallel)/2 = 13.2$ for the parallel orientation at the interface between the LC and the overlying aqueous phase because the microdroplet interface exhibits homeotropic anchoring.

Ejection of microdroplets upon the convection or the arrival of motile bacteria (Figs. 3 and 4f–h). In the system, the net force is also expressed by equation (13). Microdroplets with cationic DTAB ($C_{DTAB} = 2$ mM) have an upward F_{edl} but are elastically sequestered because F_{edl} cannot overcome F_{e1} under planar anchoring at the interface between the LC and the overlying aqueous phase (solid lines in Fig. 3e, f); $\varepsilon = (\varepsilon_\perp + \varepsilon_\parallel)/2 = 13.2$, $B = 3/4$, $F_{edl} = 6.5$ pN and $F_{e1} = -9.3$ pN at $h_1 = 1\,\mu$m for $R = 1.5\,\mu$m. The introduction of convection or of the motile bacteria in the overlying phase, however, generates a dynamic fluctuation of the director orientation at the LC interface (Figs. 3a–d and 4i), leading to changes in both F_{e1} and F_{edl}. If homeotropic anchoring is induced, for example, ε increases to 19.7 ($\varepsilon = \varepsilon_\parallel$, $F_{edl} = 9.9$ pN) from 13.2 ($F_{edl} = 6.5$ pN), while B decreases to 1/2 ($F_{e1} = -6.2$ pN) from 3/4 ($F_{e1} = -9.3$ pN). Consequently, an enhanced F_{edl} can override the reduced F_{e1}, enabling the ejection of microdroplets from the LC (dashed lines in Fig. 3e, f).

Continuous release of microdroplets through thermal tuning of the elastic properties of the LC (Extended Data Fig. 5). To illustrate this point, we used a nematic LC called E7, with $\rho_{E7} > \rho_{aq}$ and $T_{NI}^{E7} = 60\,°C$. At 25 °C, aqueous microdroplets ($0.5\,\mu m \leq R \leq 4\,\mu m$) were elastically sequestered in E7 because $F_{e1}/F_b = 1{,}194$. Both thermal tuning of the elastic properties (at $T < T_{NI}$; Extended Data Figs. 5 and 7) and isothermal tuning of the elastic properties by partitioning solutes into the LC (Extended Data Fig. 6) led to continuous release of microdroplets sequestered in the LC into the overlying aqueous phase. By contrast, when exposed to the same thermal stimulus, 5CB showed pulsatile release of microdroplets (Extended Data Fig. 5e–i). The experiment and modelling revealed that release from E7 occurred when $|F_{e1}|$ decreased below $|F_b|$ (Extended Data Fig. 5k), which in turn depended on R, h and T (thermal) or C (solute). With E7 at $T = 59\,°C$, we calculated this constraint to be satisfied for $R > 22\,\mu$m (Extended Data Fig. 5l, m). Consistent with this prediction, we observed that individual LC microdroplets with $R < 10\,\mu$m were not released. Large microdroplets with $R > 22\,\mu$m, or clusters of LC microdroplets formed through LC-mediated elastic interactions[10–14] with an effective radius of $R > 22\,\mu$m, were released (Extended Data Fig. 7).

The net force F_b^{net} acting on a guest microdroplet in a nematic phase comprises F_b and F_e (Extended Data Fig. 5k), and is calculated as

$$F_b^{net}(z) = \left[-A^2 B\pi K\left(\frac{R}{h_1+R}\right)^4\right]_{F_{e1}} + \left[A^2 B\pi K\left(\frac{R}{h_2+R}\right)^4\right]_{F_{e3}} + \left[\frac{4}{3}\pi R^3 g(\rho_{LC} - \rho_{aq})\right]_{F_b} \quad (14)$$

where $A = 2.04$ and $B = 3/4$. Here, the upward forces are F_b ($\rho_{E7} > \rho_{aq}$) and the elastic repulsion from the LC–glass interface (F_{e3}).

Parameters used in the calculations. For 5CB, at $T = 25\,°C$ we used $K_{5CB} = 7.3$ pN (ref. [31]) and $\rho_{5CB} = 1.010$ g cm^{-3} (ref. [33]), and at $T = 35\,°C$ ($= T_{NI}^{5CB}$) we used $K_{5CB} = 3$ pN (ref. [31]), $\rho_{5CB} = 1.000$ g cm^{-3} (ref. [33]), $W = 10^{-6}$ J m^{-2} (ref. [8]) and $\eta_{5CB} = 0.015$ kg m^{-1} s^{-1} (ref. [40]). The σ_{NI} value of 5CB is 1.5×10^{-5} J m^{-2} (ref. [36]). In our calculation, however, we used $\sigma_{NI} = 1.5 \times 10^{-6}$ J m^{-2} because the surfactants added to the microdroplets reduce the surface tension[41–43]. For E7, we used $K_{E7} = 14.4$ pN (25 °C), 10.25 pN (40 °C), 7 pN (50 °C) and 2 pN (59 °C)[32], and we measured $\rho_{E7} = 1.057$ g cm^{-3} (25 °C), 1.045 g cm^{-3} (40 °C), 1.037 g cm^{-3} (50 °C) and 1.028 g cm^{-3} (59 °C). We chose α and β on the basis of the experimental results as $\alpha = 9.9$ and $\beta = 0$ for $R < K/W$ and $\alpha = 0$, $\beta = 4.4$ for $R > K/W$. We used $h_1 = h_2 = 1$ mm for Fig. 1l and Extended Data Fig. 4, and $h_1 + h_2 + 2R = 3.5$ mm (thickness of LC layer in mini-wells) for Extended Data Fig. 5. For red-dye microdroplets (Fig. 1l, m and Extended Data Fig. 4), we measured $\rho_{aq} = 1.018$ g cm^{-3} (25 °C) and 1.013 g cm^{-3} (35 °C). For green-dye microdroplets (Extended Data Fig. 5l, m), we measured $\rho_{aq} = 1.012$ g cm^{-3} (25 °C), 1.004 g cm^{-3} (40 °C), 0.996 g cm^{-3} (50 °C) and 0.987 g cm^{-3} (59 °C). In F_{edl} calculations, we used $\zeta = +108$ ($C_{DTAB} = 10$ mM), +90 ($C_{DTAB} = 2$ mM), −116 ($C_{SDS} = 9$ mM), −108.5 ($C_{SDS} = 5$ mM) and −42.5 mV (no surfactants).

Materials. Nematic liquid crystals, 5CB and E7, and the chiral dopant S-811 were purchased from HCCH (Jiangsu Hecheng Display Technology Co., Ltd). Water-soluble dyes that were used as tracers were purchased from MontBlanc. SDS, DTAB, dimethyloctadecyl[3-(trimethoxysilyl)propyl]ammonium chloride (DMOAP), FITC-dextran, silver acetate and lipopolysaccharides (LPS) were purchased form Sigma-Aldrich. Lysogeny broth was purchased from Becton,

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Dickinson and Company. Transmission electron microscopy (TEM) grids were purchased from Electron Microscopy Sciences. The polymeric alignment layer (PI2555) was purchased from HD Microsystems. A Sylgard 184 silicone elastomer kit for preparing polydimethylsiloxane (PDMS) was purchased from Dow Corning. Biopsy punches were obtained from Integra Miltex.

Preparation of LCs containing aqueous microdroplets. To stabilize the dispersions of aqueous microdroplets in the LCs, we first added either SDS or DTAB at a specified concentration to the aqueous solutions of water-soluble dyes (10 wt% with respect to aqueous solution). The aqueous solutions of dyes were emulsified into the nematic LCs (5CB or E7) by vortexing for 1 min at 3,000 revolutions per minute (r.p.m.) and sonication for 10 min. The volume fractions (v%) of aqueous microdroplets (C_{aq}) dispersed in the LCs were 20 v% for Figs. 1c–h and 2g ($C_{SDS}=9$ mM), 0.5 v% for Fig. 1j–k ($C_{SDS}=9$ mM), 5 v% for Fig. 2a–c ($C_{SDS}=9$ mM), 10 v% for Fig. 3g, h and 4f–h ($C_{DTAB}=2$ mM) and 10 v% for Fig. 4b–d ($C_{SDS}=2$ mM). Each surfactant was present at a concentration below its critical micelle concentration.

Preparation of LC-filled mini-wells. Mini-wells were made of PDMS. Elastomer base and curing agent from a Sylgard elastomer kit were mixed in a ratio of 10:1. The mixture was then cured at 60 °C for 2 h. A cured PDMS disk with a diameter of 6 mm was obtained using a 6-mm biopsy punch. Subsequently, a cylindrical hole with a diameter of 3 mm was punched at the centre of the 6-mm-diameter disk using a second biopsy punch. The PDMS was treated with an oxygen plasma for 20 s and bonded to a glass substrate to create a mini-well with a depth of 3.5 mm. After fabrication, the mini-wells were stored for at least 3 days before filling with 18 μl of LCs containing guest microdroplets. Subsequently, the mini-wells were submerged into glass vials filled with 2 ml of aqueous solutions. If used within 3 days, the PDMS surface was sufficiently hydrophilic to allow water to spread between the LC phase and the PDMS surface.

Preparation of samples for microscopic observations. Experimental cells for microscopic observations were assembled from glass plates coated with a polyimide film (PI2555; Fig. 1j, k and Extended Data Fig. 2) or DMOAP (Fig. 4f–h and Extended Data Figs. 9b–e and 10), which caused planar and homeotropic alignment, respectively. PI2555 substrates were rubbed to achieve unidirectional alignment of $\boldsymbol{n}$ and were assembled in an anti-parallel fashion. The gaps (100–300 μm) between the plates were set by using double-sided tape. PI2555 cells were filled with the LCs containing microdroplets and then observed while either heating or cooling one edge of the optical cell to drive the N–I interface across the field of view. DMOAP cells were filled with aqueous solutions, followed by the LCs, and were then characterized using microscopy.

Preparation of LC films. As described in Figs. 2a–c and 3a–d (and Extended Data Figs. 3, 9f–i, 12), TEM grids were placed onto DMOAP-coated glass substrates and filled with LC containing the aqueous microdroplets. Subsequently, the films (40 μm in thickness) were submerged into aqueous baths. The DMOAP-coated glass was used to orient the LC perpendicular to the glass substrate and prevent penetration of the aqueous phase between the LC and the glass substrate.

Preparation of interfacial shear stresses (Fig. 3a–d, g and h). As shown in Figs. 3a–d, the LC films were submerged into aqueous SDS ($C_{SDS}=2$ mM; homeotropic anchoring at LC–aqueous interface, Fig. 3a) and pure water (planar anchoring at LC–aqueous interface, Fig. 3c). Subsequently, shear stresses were induced at the interface between the LC and the bulk aqueous phase by circulating the latter with a pipette. Figure 3g and h shows a mini-well filled with 5CB containing microdroplets being submerged into a water bath. Subsequently, shear stresses were induced at the interface between the LCs and the bulk aqueous phase by rotating a magnetic bar (700–800 r.p.m.).

Preparation of cholesteric LC (Fig. 4a–d). 20 wt% of chiral dopant (S-811) was dissolved in 5CB. The clearing temperature of the LC was measured to be $T_{NI}=27$ °C.

Preparation of bacterial dispersions (Fig. 4f–h). *Escherichia coli* (strain MG1655) were grown aerobically in 1 ml of lysogeny broth [1% (w/v) tryptone, 0.5% (w/v) yeast extract and 1% (w/v) NaCl] at temperature $T=37$ °C with agitation (200 r.p.m.) for 12 h. To achieve motile bacteria, the culture was diluted into 2 ml of fresh lysogeny broth in a 1:100 ratio and the bacteria were grown again for 2 h ($T=37$ °C, 200 r.p.m.). The density of bacteria in the resulting dispersion was 10^7–10^8 cells ml^{-1}.

Optical responses in Figs. 1e and 4i. The optical response shown in Fig. 1e is the intensity measured from the micrographs of mini-wells between crossed polarizers (insets in Fig. 1e). For the optical response in Fig. 4i, the transmittance at the LC–aqueous interface was averaged for every frame and the average values are plotted for 4 s for each sequence (Fig. 4f–h).

LC system with multiple stimuli (Extended Data Fig. 11). In this system, the mechanisms leading to the release of microcargo are the same as those described in Fig. 1 (thermal stimulus) and 2 (chemical stimulus). Specifically, the net force acting on microdroplets in the LC is described by $F_{edl}+F^{net}_{h,N}$ (see equations (3) and (11)) during heating and $F_{edl}+F^{net}_{c,NI}$ (see equation (7)) during cooling. Because the sign and amplitude of F_{edl} can be manipulated by charged surfactants added into overlying aqueous phases, we predicted that the combination of thermal and chemical stimuli would enable selective release of two agents that are oppositely charged. To test this prediction, Wells 1 and 2 were filled with 5CB containing microdroplets ($C_{aq}=20$ v%) doped with DTAB ($C_{DTAB}=9$ mM, $\zeta>0$, and green tracer) and SDS ($C_{SDS}=9$ mM, $\zeta<0$, and red tracer), respectively, and then were submerged into an aqueous SDS solution (3 mM, $\zeta<0$). Under these conditions, $F_{edl}>0$ in Well 1 ($\zeta<0$ at the interface between the LCs and the overlying aqueous phase, but $\zeta>0$ at their interface with the microdroplets), while $F_{edl}<0$ in Well 2 ($\zeta<0$ at both interfaces). Upon thermally triggering phase transitions (0th to 4th phase transitions, $T_h=50$ °C and $T_c=25$ °C), Well 1 released green tracer owing to the attractive F_{edl} and the elastic forces (F_{e2} for heating, equation (3); elastic strain F_{e*} and interfacial tension F_{it} forces for cooling, equation (7)) generated by the motion of the N–I interface. By contrast, no release was observed from Well 2 because the attractive elastic forces did not override the repulsive F_{edl}. After four phase transitions, aqueous DTAB ($C_{DTAB}=3$ mM) was introduced into the bath to reverse the charge at the interface between the LCs and the overlying aqueous phase from $\zeta<0$ to $\zeta>0$ (and thus $F_{edl}<0$ in Well 1 and $F_{edl}>0$ in Well 2). Consequently, the elastic forces were able to trigger the release of microcargo from Well 2 ($F_{edl}>0$, red tracer) but not Well 1 ($F_{edl}<0$, green tracer).

Temperature control. The temperature was controlled using a STC200 hot stage and a controller (Instec Inc.) with and accuracy of 0.1 °C. Both heating and cooling were achieved by circulation of cold water. The rate of temperature change was typically ± 15 °C min^{-1}.

Measurement for the mass of tracer released. 6 μl of aqueous solution was collected from the baths contacting the LCs after each N–I phase transition (Fig. 1 and Extended Data Fig. 1i and 11f) or every 2–5 min (Fig. 2g and Extended Data Fig. 5j). Prior to collection of a sample, the baths were gently agitated to uniformly mix the tracer released from the LCs through the overlying aqueous solution. We estimated the mass of tracer released from the absorbance of the samples at a wavelength corresponding to peak tracer absorbance, which was measured using a NanoDrop 2000 (Thermo Scientific) spectrophotometer.

Zeta potential measurement. 5CB (0.01 v% ≥ C_{5CB} ≥ 0.001 v%) was emulsified in aqueous solution (water or aqueous solutions of SDS or DTAB) using a homogenizer. The zeta potentials on the aqueous side of the LC–aqueous interface were measured using a Zetasizer Nano instrument (Malvern Instruments Ltd).

Additional observations on the transport of microdroplets by propagating N–I interfaces. We made two additional observations in the experiments shown in Fig. 1j, k and Extended Data Fig. 2. First, we observed single microdroplets or microdroplet clusters with $R<R^*$ to be transported initially by the moving N–I interface (denoted by dotted circles in Extended Data Fig. 2c, d, g and h). However, as the moving interface formed bigger clusters with $R>R^*$ by collecting additional microdroplets, we observed some microdroplets from the cluster to be left behind the interface, as illustrated in Extended Data Fig. 2i–l. West et al.[17] observed similar behaviours with solid particles and attributed them to an increase in the effective radius of the particles due to aggregation. Importantly, this observation provides insight into why only a fraction of the microdroplets was released at each phase transition (Fig. 1g).

Second, we observed microdroplets to occasionally coalesce upon heating (denoted by white arrows in Extended Data Fig. 2c, d). Consequently, large microdroplets formed, which were observed to remain behind the moving N–I interface. This latter observation provides insight into why the amount of tracer released after 20 cycles corresponded to approximately 40% of the tracer loaded initially into the 5CB (Fig. 1i). Overall, these results indicate that the fraction of microdroplets released can be manipulated by tuning the clustering size and coalescence of microdroplets.

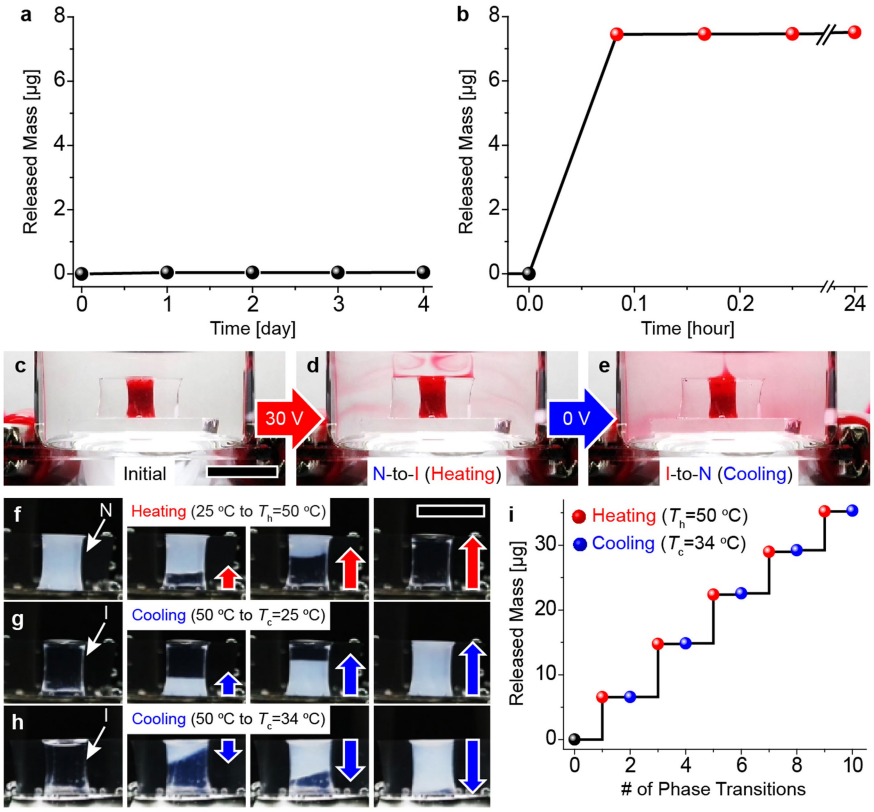
Data availability. The authors declare that the main data supporting the findings of this study are available within the paper and in Supplementary Information. Additional data are available from the corresponding author upon reasonable request.

31. Bogi, A. & Faetti, S. Elastic, dielectric and optical constants of 4′-pentyl-4-cyanobiphenyl. *Liq. Cryst.* **28**, 729–739 (2001).
32. Raynes, E. P., Tough, R. J. A. & Davies, K. A. Voltage dependence of the capacitance of a twisted nematic liquid crystal layer. *Mol. Cryst. Liq. Cryst.* **56**, 63–68 (1979).
33. Kim, J.-W., Kim, H., Lee, M. & Magda, J. J. Interfacial tension of a nematic liquid crystal/water interface with homeotropic surface alignment. *Langmuir* **20**, 8110–8113 (2004)
34. Israelachvili, J. N. *Intermolecular and Surface Forces* 3rd edn (Elsevier Science, Burlington, 2010).
35. Zimmermann, N., Junnemann-Held, G., Collings, P. J. & Kitzerow, H.-S. Self-organized assemblies of colloidal particles obtained from an aligned chromonic liquid crystal dispersion. *Soft Matter* **11**, 1547–1553 (2015).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

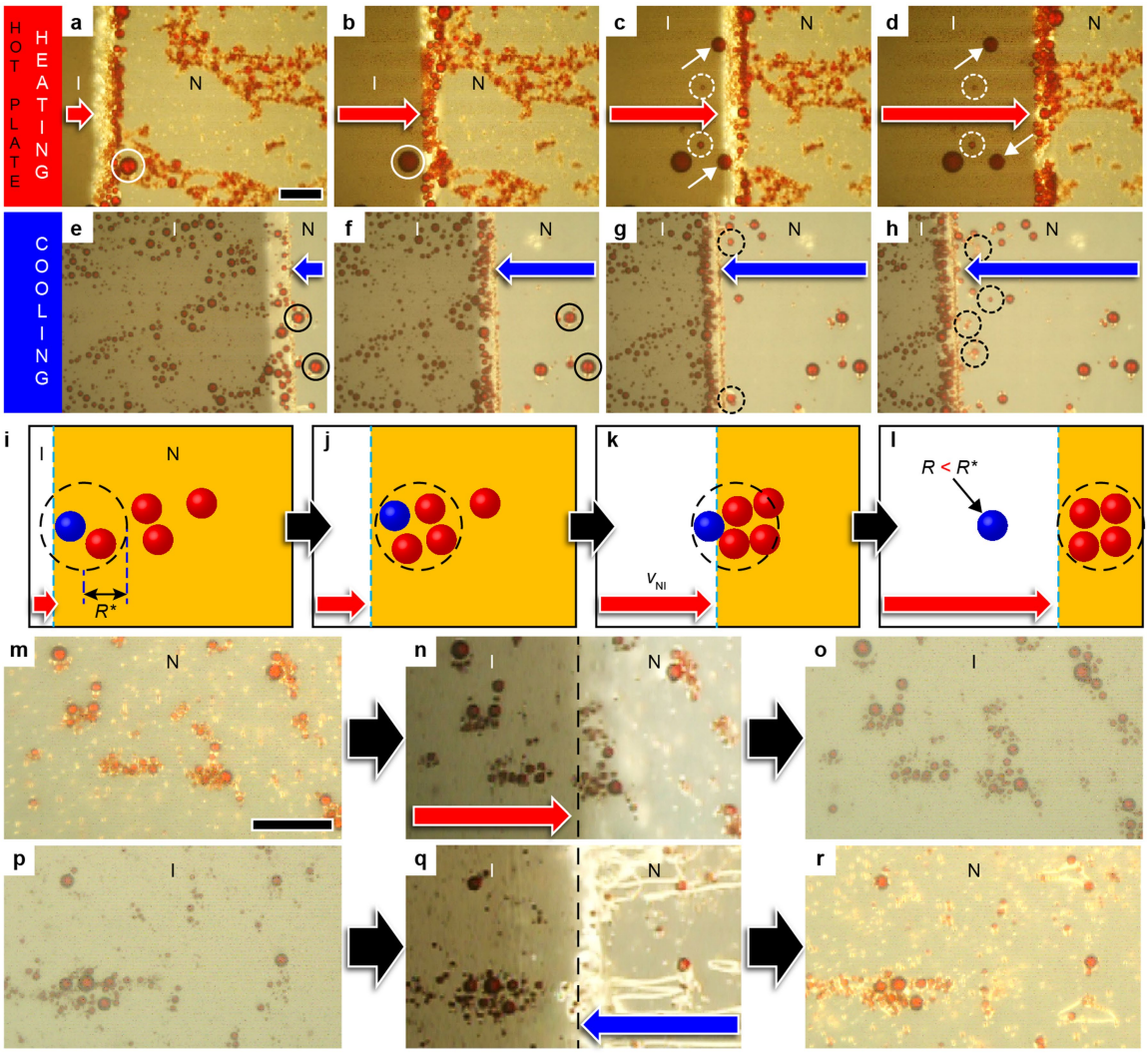
36. Faetti, S. & Palleschi, V. Measurements of the interfacial tension between nematic and isotropic phase of some cyanobiphenyls. *J. Chem. Phys.* **81**, 6254–6258 (1984).
37. Andrienko, D., Tasinkevych, M., Patricio, P. & da Gama, M. M. T. Interaction of colloids with a nematic-isotropic interface. *Phys. Rev. E* **69**, 021706 (2004).
38. Shah, R. R. & Abbott, N. L. Coupling of the orientations of liquid crystals to electrical double layers formed by the dissociation of surface-immobilized salts. *J. Phys. Chem. B* **105**, 4936–4950 (2001).
39. Brown, M. A. et al. Determination of surface potential and electrical double-layer structure at the aqueous electrolyte-nanoparticle interface. *Phys. Rev. X* **6**, 011007 (2016).
40. Janik, J., Krol-Otwinowska, A., Sokolowska, D. & Moscicki, J. K. Pendulum viscometer: a new method for measurement of Miesowicz nematic shear viscosity coefficients η_1 and η_2. *Rev. Sci. Instrum.* **77**, 123906 (2006).
41. Holmberg, K., Jönsson, B., Kronberg, B. & Lindman, B. *Surfactants and Polymer in Aqueous Solution* (John Wiley & Sons, Chichester, 2002).
42. Harth, K., Shepherd, L. M., Honaker, J. & Stannarius, R. Dynamic interface tension of a smectic liquid crystal in anionic surfactant solutions. *Phys. Chem. Chem. Phys.* **17**, 26198–26206 (2015).
43. Ong, L. H. & Yang, K.-L. Surfactant-driven assembly of poly(ethylenimine)-coated microparticles at the liquid crystal/water interface. *J. Phys. Chem. B* **120**, 825–833 (2016).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



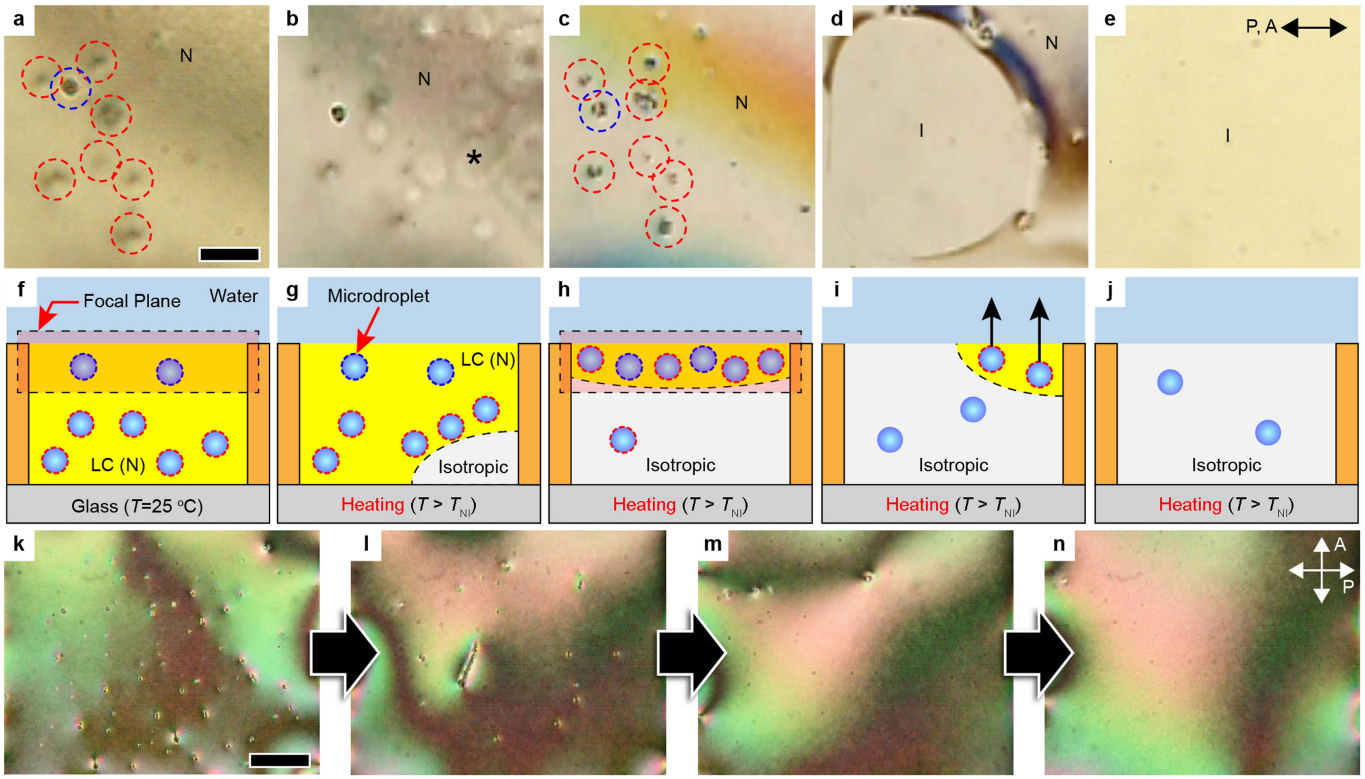
Extended Data Fig. 1 | Thermally triggered ejection of microdroplets. **a**, **b**, Mass of tracer released from LCs (5CB) as a function of time, before (**a**) and after (**b**) an N-to-I phase transition (corresponds to Fig. 1). Black and red points indicate the mass released before and after an N–I phase transition, respectively. **c–e**, Sequential photographs of the release of microdroplets (C_{SDS} = 9 mM and red tracer) from the LC, which is triggered by phase transitions induced by resistive heating. We used C_{aq} = 20 v% and 30 V for heating (to T_h = 60 °C, **d**) and 0 V for cooling (to room temperature, **e**). Heating of the sample from below was achieved by passing a current through an indium-tin-oxide electrode coated on glass. The motion of the N–I interface was upward for both heating and cooling. **f–h**, Dependence of the direction of motion of the N–I interface on T_h and T_c. **f**, Optical images showing a propagation of the N–I interface across a mini-well filled with 5CB (containing no microdroplets) in an aqueous bath upon heating from T = 25 °C to T_h = 50 °C (N-to-I phase transition). Upon heating the sample from below, the interface moved upwards (towards the interface between the LCs and the overlying aqueous phase) regardless of T_h. **g**, Upward motion of the N–I interface upon cooling from T = 50 °C to T_c = 25 °C (I-to-N phase transition). **h**, Downward motion of the N–I interface upon cooling from T = 50 °C to T_c = 34 °C (I-to-N phase transition). Scale bars, 5 mm. **i**, Released mass of tracer as a function of the number of phase transitions with T_h = 50 °C and T_c = 34 °C; C_{aq} = 20 v% (C_{SDS} = 9 mM).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



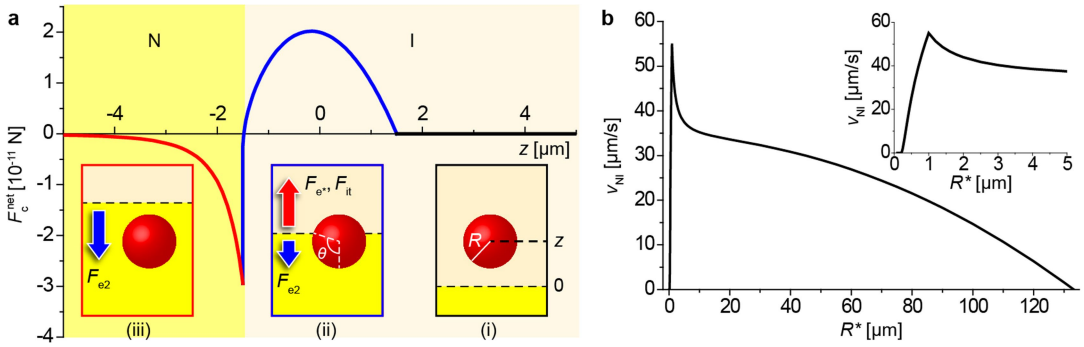
Extended Data Fig. 2 | Behaviours of microdroplet clusters during passage of the N–I interface. a–h, Sequential micrographs showing the behaviours of microdroplet clusters dispersed in an LC (5CB) during passage of an N–I interface upon an N-to-I (heating, **a–d**) and an I-to-N (cooling, **e–h**) phase transition. We used $C_{aq} = 3$ v% ($C_{SDS} = 2$ mM) and measured $v_{NI} = 10\,\mu m\,s^{-1}$ for heating, $v_{NI} = 35\,\mu m\,s^{-1}$ for cooling and $R^* \approx 10\,\mu m$ for both cases. Scale bar, 100 μm (see Methods, 'Additional observations on the transport of microdroplets by propagating N–I interfaces' for more details). Red and blue arrows indicate the direction of motion of the N–I interface. Solid and dotted circles indicate microdroplets with $R > R^*$ ($= 10\,\mu m$) and $R < R^*$, respectively. White arrows indicate microdroplets that coalesced while being transported by the moving N–I interface. We note that microdroplets with $R < R^*$ were left behind the N–I interface in **c** because they were shed from clusters, as illustrated in **i–l**. **i–l**, Illustration of a microdroplet cluster being transported by a moving N–I interface. **i**, Single microdroplets or microdroplet clusters with $R < R^*$ are transported by a moving N–I interface. **j**, As the moving interface collects more microdroplets, the microdroplet clusters formed at the interface increase in size. **k**, When the effective radius of a microdroplet cluster exceeds R^*, the interface no longer transports the cluster. **l**, Because some of microdroplets from the cluster are left behind the N–I interface, the cluster becomes smaller than R^* and thus is transported again by the interface. **m–r**, Evidence that microdroplets with $R > R^*$ are not transported by an N–I interface moving at high speed ($v_{NI} = 100\,\mu m\,s^{-1}$) during N-to-I (**m–o**) and I-to-N (**p–r**) phase transitions. Here $C_{aq} = 0.5$ v% and $C_{SDS} = 2$ mM. Scale bar, 100 μm. The left (**m**, **p**) and right (**o**, **r**) columns show optical micrographs before and after passage of the N–I interface, respectively, and the middle column (**n**, **q**) shows micrographs taken during passage of the N–I interface. The positions of microdroplets before and after the passage of the N–I interface were unchanged, revealing that $R > R^*$ for the rapidly moving interface.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



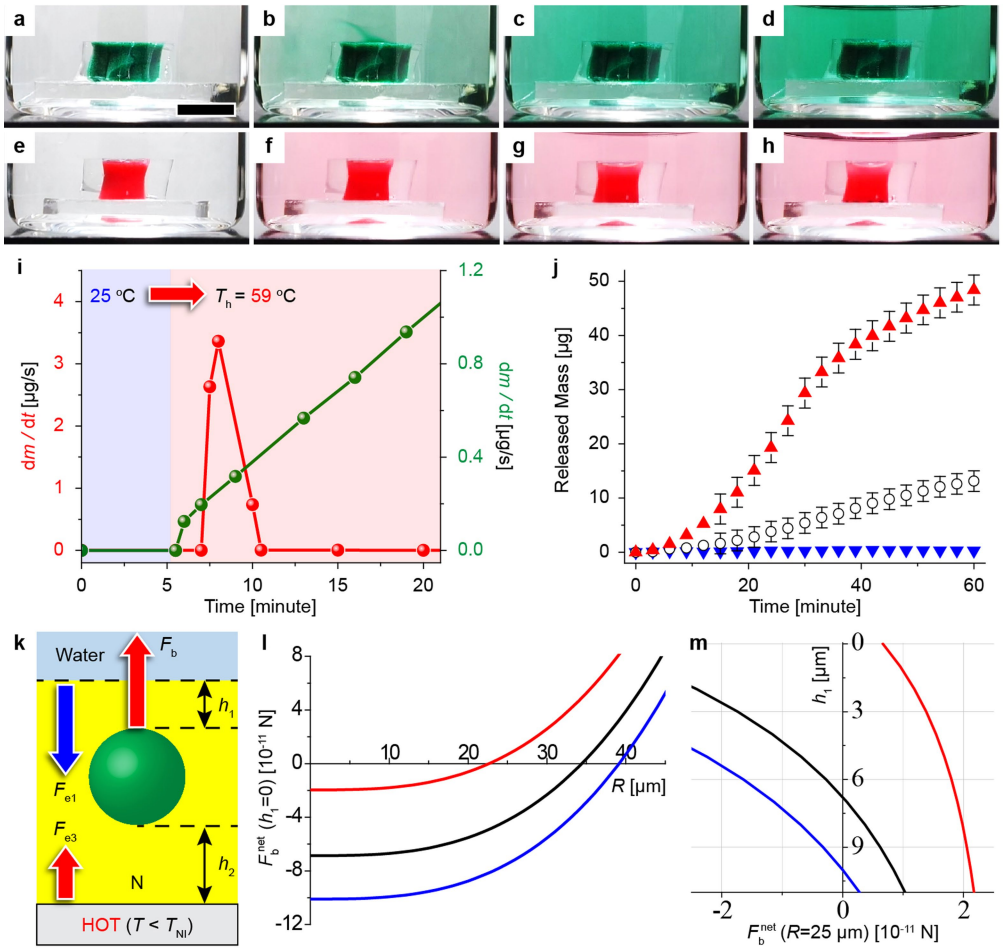
Extended Data Fig. 3 | Transport of microdroplets by an N–I interface propagating across an LC film. a–j, Sequential micrographs (**a–e**, top view) and corresponding illustrations (**f–j**, side view) of microdroplets transported by a moving N–I interface upon heating in an LC film (5CB, 40 μm in thickness). The focal plane is near the interface between the LC and the overlying water (red boxes in **f** and **h**). **a**, **f**, Microdroplets are dispersed initially in the LC bulk. **b**, **g**, When the bottom of the LC film is heated to $T_h = 50\,°C$ ($> T_{NI}$), the N-to-I transition first occurs at the LC–glass interface (denoted by the asterisk in **b**) and the N–I interface propagates upwards, towards the LC–water interface. **c**, **h**, Microdroplets that were out of focus (red dashed circles in **a** and **f**) move into focus, revealing that the moving interface transported the microdroplets towards the LC–water interface. **d**, **i**, As the N–I interface reaches the LC–water interface, the microdroplets disappear, consistent with their fusion with the overlying aqueous phase. **e**, **j**, After the phase transition, some microdroplets remain in the LC layer. **k–n**, Micrographs showing the decrease in the population of microdroplets in the LC at $T = 25\,°C$; before any phase transitions (**k**) and after 2 (**l**), 4 (**m**) and 6 (**n**) phase transitions. Scale bars, 50 μm. P and A indicate the orientations of the polarizer and analyser, respectively. $T_h = 50\,°C$, $T_c = 25\,°C$, $C_{aq} = 5$ vol% ($C_{SDS} = 9$ mM) and $v_{NI} = 8\,\mu m\,s^{-1}$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



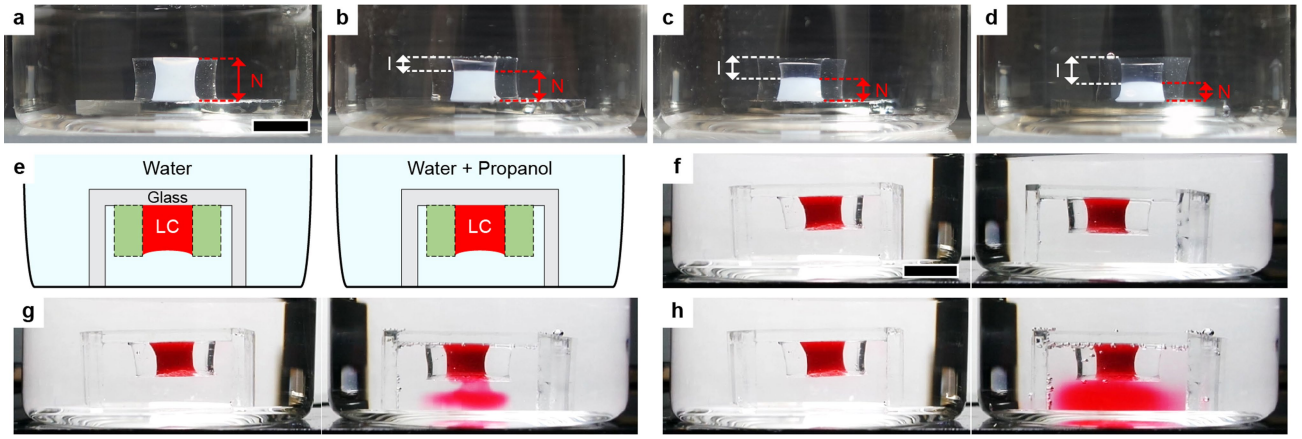
Extended Data Fig. 4 | Calculated net force $F_c^{net}(z)$ acting on a microdroplet and calculated dependence of R^* on v_{NI} during an I-to-N phase transition (cooling). a, $F_c^{net}(z)$ for a quasi-static microdroplet of $R = 1.5\,\mu m$ in 5CB. The insets show microdroplets at $z \geq R$ (black line; (i)), at $-R < z < R$ (blue line; (ii)) and at $z \leq -R$ (red line; (iii)). **b**, Critical radius R^* of a microdroplet as a function of v_{NI} upon cooling. See Methods for details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



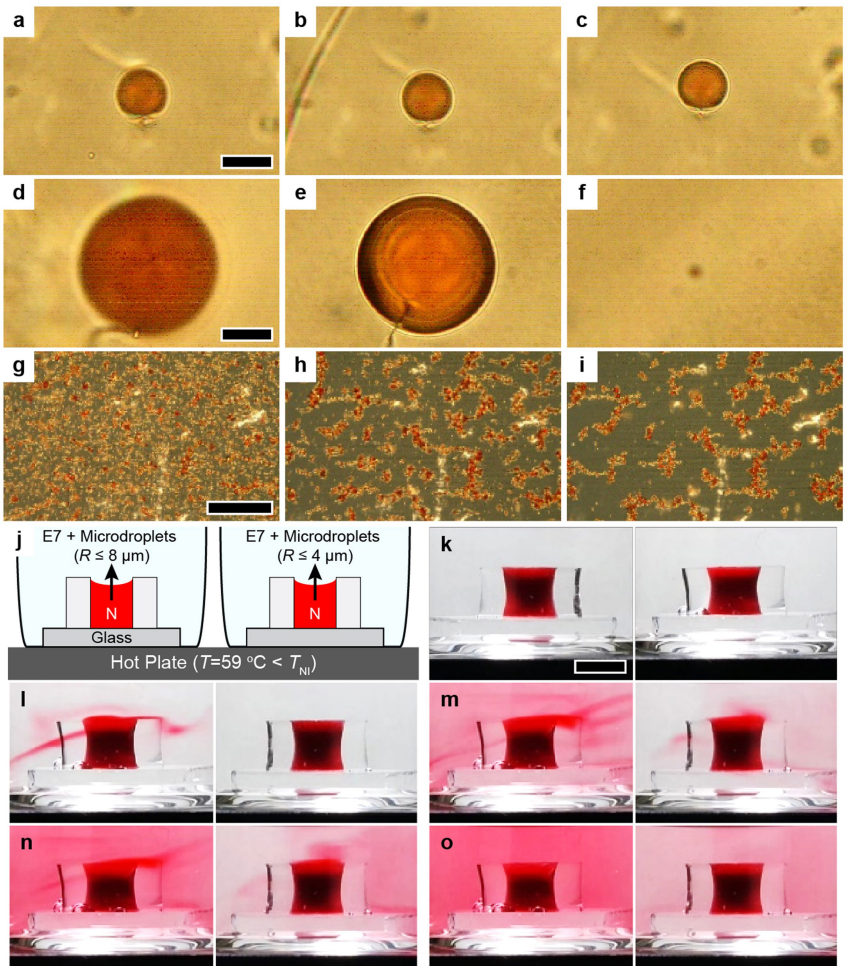
Extended Data Fig. 5 | Influence of LC phase behaviour and elastic properties on dynamics of release of dispersed microdroplets. **a–h**, Sequential photographs showing continuous release (**a–d**) of dispersed microdroplets (green tracer and $C_{SDS} = 2$ mM) from a nematic LC (E7) in response to a thermal trigger $T_h < T_{NI}^{5CB}$ ($\rho_{E7} > \rho_{aq}$), and for pulsatile release of dispersed microdroplets (red tracer and $C_{SDS} = 9$ mM) from 5CB ($\rho_{5CB} < \rho_{aq}$, $T_h > T_{NI}^{5CB}$) at 0 min (**a**, **e**), 15 min (**b**, **f**), 60 min (**c**, **g**) and 120 min (**d**, **h**) after heating of the baths from below to $T_h = 59$ °C. We used $C_{aq} = 30$ v% for **a–d** and $C_{aq} = 20$ v% for **e–h**. Scale bar, 5 mm. **i**, Corresponding time derivative of the released mass m of the tracer (dm/dt; t, time) for pulsatile (red line and circles) and continuous (green line and circles) release. **j**, Released mass of green tracers from E7 (continuous release) with respect to time at representative temperatures of $T = 40$ °C (blue triangles), 50 °C (circles) and 59 °C (red triangles). The data are mean values and the error bars are 1 s.d. ($n = 5$). **k**, Forces acting on an aqueous microdroplet in E7. **l**, **m**, Calculated net force F_b^{net} acting on a microdroplet in E7 as a function of R at $h_1 = 0$ (**l**) and as a function of h_1 for $R = 25$ μm (**m**) at $T = 40$ °C (blue line), 50 °C (black line) and 59 °C (red line). See Methods for details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



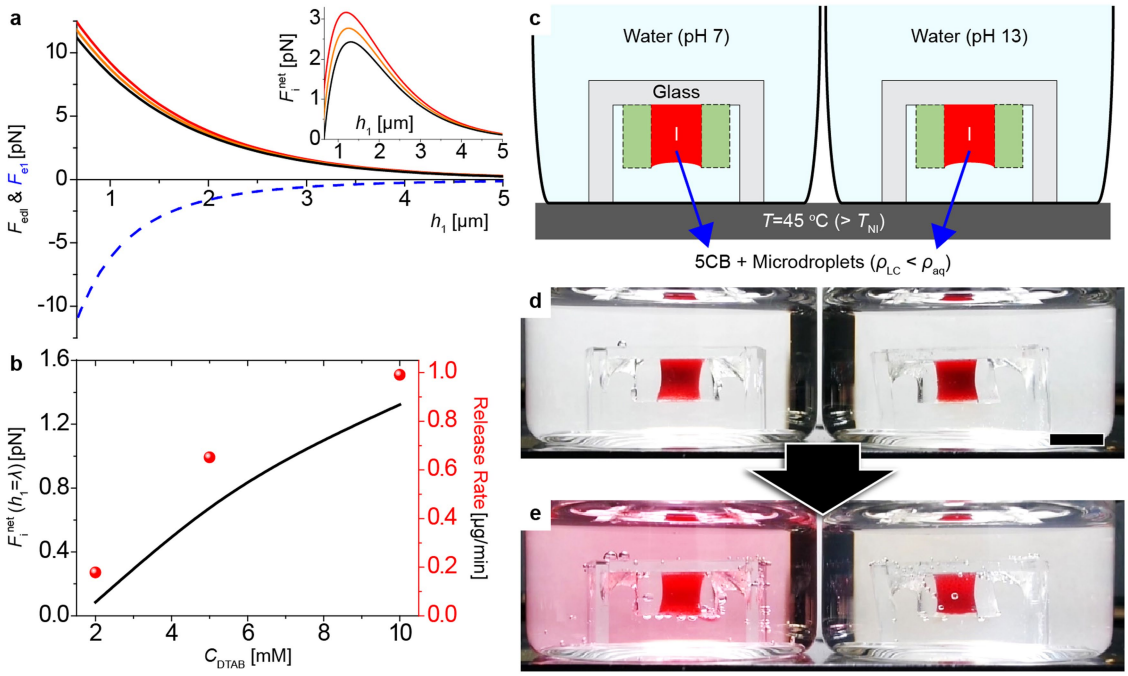
Extended Data Fig. 6 | Isothermal release of microdroplets from an LC by a solute-triggered N-to-I phase transition. a–d, Sequential photographs of a solute-triggered N-to-I phase transition of 5CB at $T = 25\,°C$ ($T_{NI}^{5CB} = 35\ °C$) at 0 h (**a**), 1 h (**b**), 2 h (**c**) and 3 h (**d**) after the addition of propanol to the overlying water. As the propanol diffused into the 5CB, an N-to-I transition occurred first at the interface between the LC and the aqueous solution and propagated into the LC bulk. **e**, Illustration of inverted mini-wells filled with 5CB containing microdroplets (red tracer), placed in baths containing pure water (left) and water containing propanol ($C_{propanol} = 16$ v%; right). **f–h**, Sequential photographs of the mini-wells at 0 min (**f**), 5 min (**g**) and 30 min (**h**) after the mini-wells were submerged into the baths; $C_{aq} = 10$ v% ($C_{SDS} = 9$ mM). Although F_b ($\rho_{LC} < \rho_{aq}$) promotes the release of tracers, no release of red tracers was observed in the left bath owing to strong elastic sequestration. In the right bath, however, the red tracers were continuously released as the elastic barrier was removed by the solute-induced N-to-I phase transition. Scale bars, 5 mm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



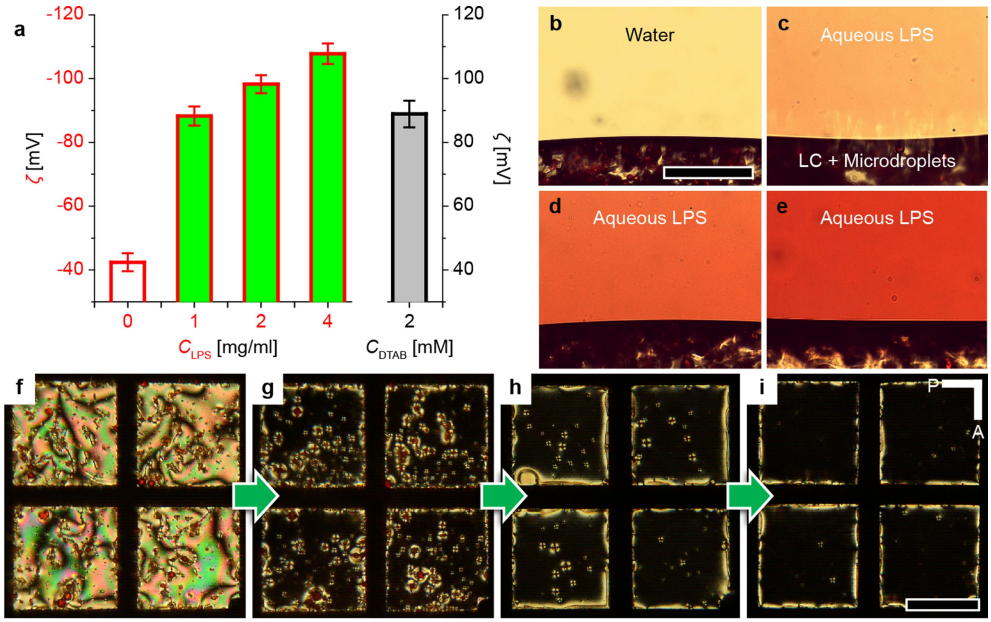
Extended Data Fig. 7 | Influence of the size and clustering of microdroplets on release from the LC. a–f, Optical micrographs of microdroplets ($C_{SDS} = 2$ mM and red tracer), with $R = 9.5\,\mu$m (**a–c**) and 27 μm (**d–f**) in an LC (E7) at 25 °C (**a**, **d**), 50 °C (**b**, **e**) and 59 °C (**c**, **f**); $\rho_{E7} > \rho_{aq}$ and $C_{aq} = 1$ v%. Scale bars, 20 μm. The microdroplets were elastically trapped in the nematic LC bulk at 25 °C. As the temperature increased to 50 °C ($R > 34\,\mu$m for release), the microdroplets moved upwards and into focus, but were not dispensed into the overlying water; the focal plane was near the interface between the LC and the overlying aqueous solution. At 59 °C ($R > 23\,\mu$m for release), we observed the larger microdroplet ($R = 27\,\mu$m) to escape into the overlying aqueous phase, whereas the smaller microdroplet ($R = 9.5\,\mu$m) remained elastically trapped in the nematic bulk. This observation is in good agreement with our theoretical prediction (Extended Data Fig. 5l, m). **g–i**, Micrographs of the clustering of microdroplets at 0 min (**g**), 30 min (**h**) and 180 min (**i**) after they were dispersed in the LC; $C_{aq} = 2$ v% ($C_{SDS} = 9$ mM). Scale bar, 200 μm. **j–o**, Illustration (**j**) and sequential micrographs (**k–o**) of thermally triggered, continuous release of microdroplets ($C_{SDS} = 9$ mM and red tracer) from mini-wells filled with E7 containing microdroplets of different sizes, at 0 min (**k**), 7 min (**l**), 10 min (**m**), 12 min (**n**) and 15 min (**o**) after the baths were heated to $T_h = 59$ °C($< T_{NI}$); $\rho_{E7} > \rho_{aq}$ ($F_b > 0$) and $C_{aq} = 20$ v%. Scale bar, 5 mm. The mini-well containing the larger microdroplets (left bath) exhibited a higher release rate due to the facile formation of microdroplet clusters with a radius higher than that for which $F_b^{net} > 0$, consistent with the theoretical model whose results are shown in Extended Data Fig. 5l, m.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



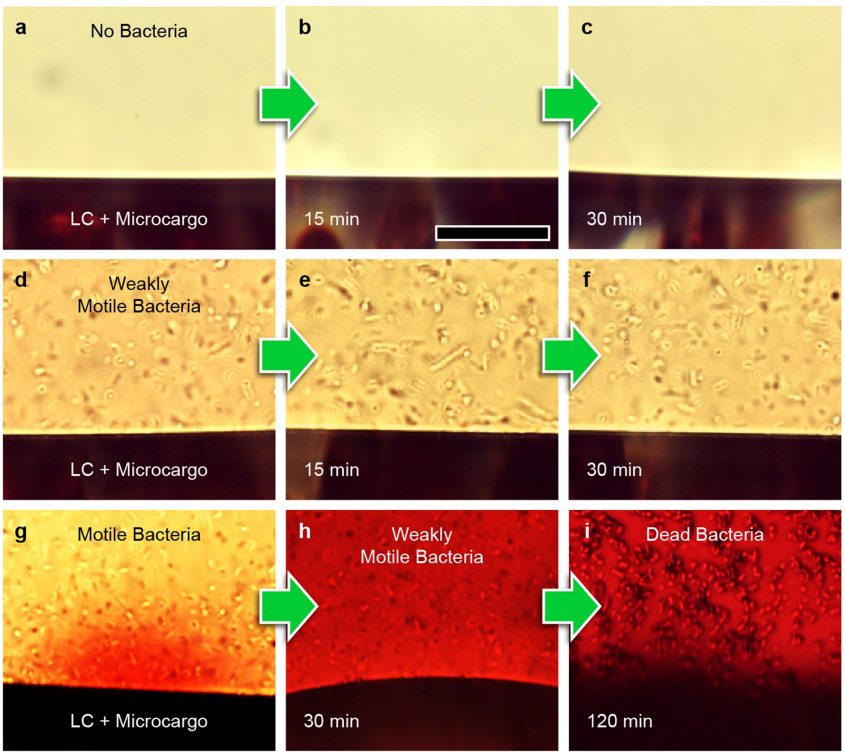
Extended Data Fig. 8 | Role of electrical double-layer interactions in the release of microdroplets from an LC. a, Calculated repulsive elastic (F_{e1}, dashed line) and attractive electrical double-layer (F_{edl}, solid lines) forces acting on a microdroplet ($C_{SDS} = 9$ mM, $R = 1.5$ μm) as a function of h_1 following the addition of DTAB with $C_{DTAB} = 10$ mM (red), 5 mM (orange) and 2 mM (black) to the overlying aqueous phase. The inset shows the corresponding net forces ($F_i^{net} = F_{e1} + F_{edl}$). **b**, F_i^{net} near the interface between the LC and the overlying aqueous phase ($h_1 = \lambda$), and initial release rate of microdroplets (from 0 to 60 min in Fig. 2g) as a function of C_{DTAB} added into the overlying aqueous phase. **c**, Illustration of inverted mini-wells in baths with water (pH 7, left bath) and alkaline water (pH 13, right bath) at $T = 45$ °C ($> T_{NI}$). **d**, **e**, Sequential photographs of mini-wells filled with 5CB containing aqueous microdroplets (red tracer) at 0 min (**b**) and 60 min (**c**) after an N-to-I phase transition, with $C_{aq} = 10$ v% ($C_{SDS} = 9$ mM) and $T_h = 50$ °C. Scale bar, 5 mm. Because $\rho_{LC} < \rho_{aq}$ ($F_b < 0$), tracers are continuously released from an isotropic phase of 5CB ($F_{e1} = 0$) into the pure water (left bath). In the alkaline water (right bath), however, the release is suppressed because of the introduction of repulsive charge interactions between the LC interface in alkaline water (negatively charged) and SDS-containing aqueous microdroplets (negatively charged).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



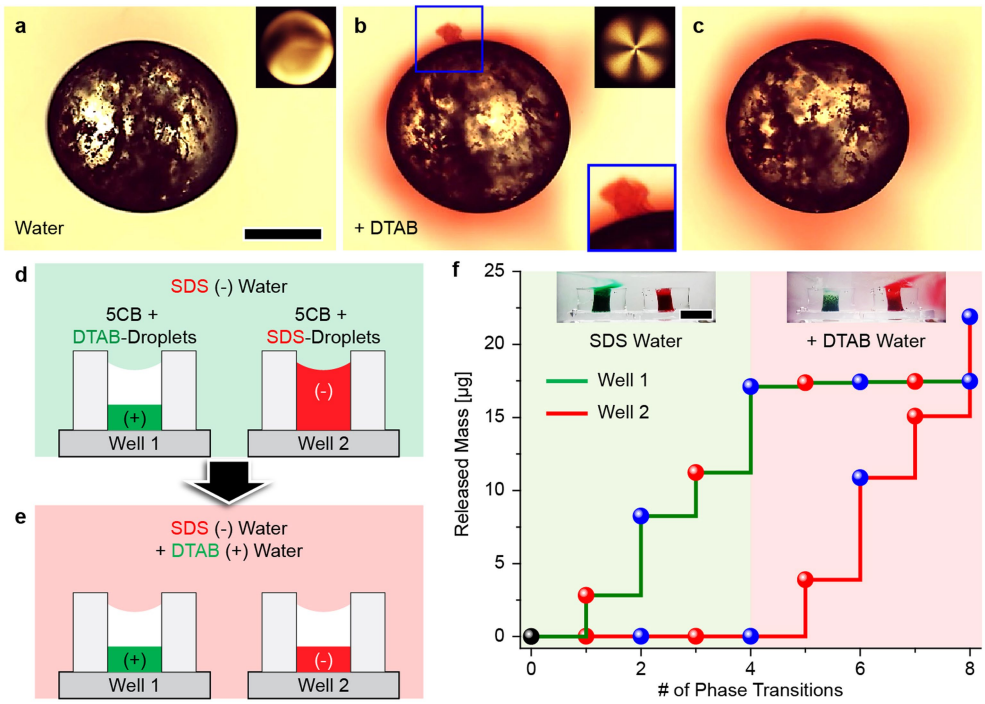
Extended Data Fig. 9 | Triggering of the release of dispersed microdroplets by interfacial charge interactions of biological molecules. a, Zeta potential ζ at LC–aqueous interface without (white bar) and with lipopolysaccharides (LPS; green bars) from *Escherichia coli*, or with DTAB (grey bar). The data are mean values and the error bars are 1 s.d. ($n > 3$). **b–e**, Micrographs (side view) showing the ejection of microdroplets ($C_{DTAB} = 2$ mM and red tracer) from LC (5CB) 30 min after addition of $C_{LPS} = 0$ mg ml^{-1} (**b**), 1 mg ml^{-1} (**c**), 2 mg ml^{-1} (**d**) and 4 mg ml^{-1} (**e**) to the overlying aqueous phase. **f–i**, Sequential micrographs (top view) showing the ejection of microdroplets before (**f**) and at 0 min (**g**), 15 min (**h**) and 30 min (**i**) after the addition of $C_{LPS} = 4$ mg ml^{-1} into the overlying aqueous phase. Scale bars, 200 μm. P and A indicate the orientations of the polarizer and analyser, respectively. In the presence of LPS, microdroplets are ejected continuously from the LC, as evidenced by the release of red tracer (**b–e**) and the decrease in the population of aqueous microdroplets within the thin LC film (40 μm in thickness; **f–i**). The release rate is enhanced with an increase in C_{LPS} (**b–e**), consistent with release controlled by interfacial charge interactions.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



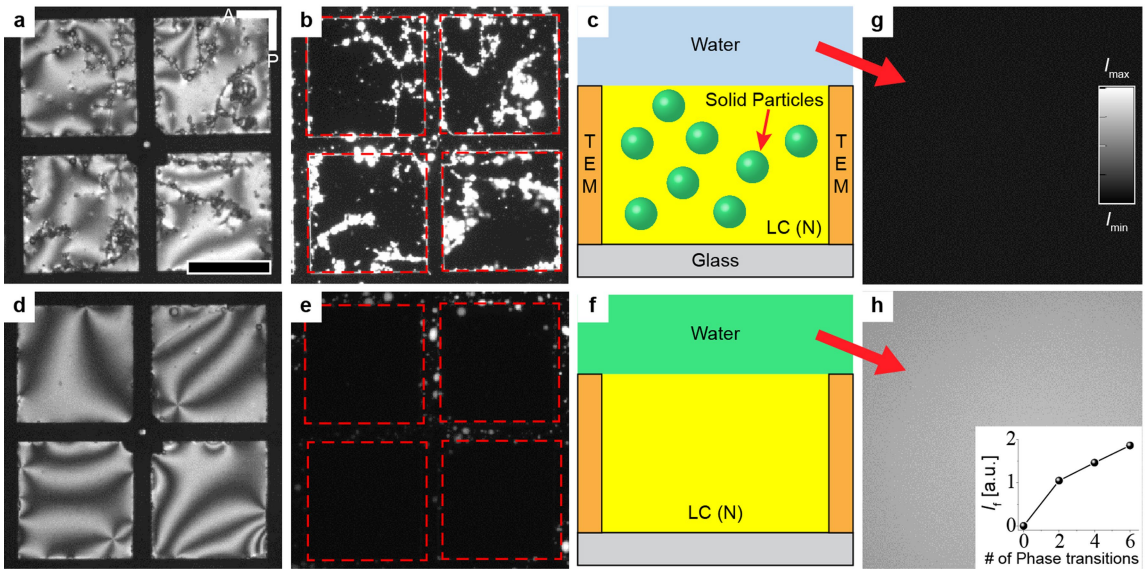
Extended Data Fig. 10 | Ejection of microdroplets from the LC by motile bacteria. a–f, Sequential micrographs showing that microdroplets containing anti-bacterial agent (2 mM DTAB and 1 wt% silver acetate) are not ejected from the LC (5CB) in the absence of bacteria (**a–c**) and at 0 min (**a**, **d**), 15 min (**b**, **e**) and 30 min (**c**, **f**) after the arrival of weakly motile bacteria (**d–f**); $C_{aq} = 10$ v%. **g–i**, Sequential micrographs showing ejection of microdroplets from the LC at 0 min (**g**), 30 min (**h**) and 120 min (**i**) after the arrival of motile bacteria and subsequent bacterial death and aggregation. Scale bar, 40 μm. **g**, Motile bacteria generate shear stresses at the LC interfaces, triggering the release of microdroplets containing anti-bacterial agent. **h**, As the anti-bacterial agents are released, the bacteria become less motile. At 30 min, we observe dead bacteria and cessation of the triggered release due to the decrease in the number of motile bacteria. The amount of silver acetate released is 1–2 $\mu g\,\mu l^{-1}$. **i**, Two hours after the arrival of motile bacteria, only dead (non-motile) bacteria are observed.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 11 | LC systems with complex geometries and responsiveness to multiple stimuli. a–c, Isothermal release of aqueous microdroplets ($C_{SDS} = 9$ mM and red tracer) from a large LC emulsion droplet (5CB) before (**a**) and at 50 s (**b**) and 80 s (**c**) after the addition of DTAB (10 mM) into surrounding aqueous phase; $C_{aq} = 5$ v%. Scale bar, 200 µm. The insets in **a** and **b** are optical micrographs (crossed polars) of the LC droplet, showing the optical response. **d–f**, Selective release of two agents, triggered by a combination of chemical and thermal stimuli (see Methods for details). **d**, In an aqueous SDS bath, the thermal stimulus triggers release of DTAB-doped microdroplets (green tracer) from Well 1, but not of SDS-doped microdroplets (red tracer) from Well 2 (1st to 4th phase transition; left inset in **f**). **e**, Addition of DTAB into the bath reverses the charge at the interface between the LCs and the overlying aqueous phase, thus enabling the ejection of microdroplets from Well 2, but not from Well 1 (5th to 8th phase transition; right inset in **f**). Red and blue points in **f** correspond to the released mass of tracers after N-to-I and I-to-N phase transitions, respectively. Scale bar, 5 mm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 12 | Triggered release of water-soluble solid microparticles from an LC film. a–f, Polarizing (**a**, **d**; top view) and fluorescence (**b**, **e**; top view) micrographs and schematic illustrations (**c**, **f**; side view) of an LC film at 25 °C in the initial N phase (**a–c**) and after six phase transitions (**d–f**) with $T_h = 50$ °C and $T_c = 25$ °C. Scale bar, 200 μm. The LC film (5CB, 40 μm in thickness) contained solid microparticles of FITC-dextran (1–2 wt%). Before any phase transition, a strong fluorescence signal was detected from the FITC-dextran microparticles that were sequestered in the LC film (**b**), but not from the water bath (**g**), indicating that the solid microparticles were trapped in the LC. After six N–I phase transitions, however, no fluorescence signal was detected from the LC (**e**), whereas the water bath showed a strong fluorescence signal (**h**). The inset in **h** shows the fluorescence intensity I_f from the bath as a function of the number of phase transitions.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0131-1

Rapid emergence of subaerial landmasses and onset of a modern hydrologic cycle 2.5 billion years ago

I. N. Bindeman[1]*, D. O. Zakharov[1], J. Palandri[1], N. D. Greber[2], N. Dauphas[3], G. J. Retallack[1], A. Hofmann[4], J. S. Lackey[5] & A. Bekker[4,6]

The history of the growth of continental crust is uncertain, and several different models that involve a gradual, decelerating, or stepwise process have been proposed[1–4]. Even more uncertain is the timing and the secular trend of the emergence of most landmasses above the sea (subaerial landmasses), with estimates ranging from about one billion to three billion years ago[5–7]. The area of emerged crust influences global climate feedbacks and the supply of nutrients to the oceans[8], and therefore connects Earth's crustal evolution to surface environmental conditions[9–11]. Here we use the triple-oxygen-isotope composition of shales from all continents, spanning 3.7 billion years, to provide constraints on the emergence of continents over time. Our measurements show a stepwise total decrease of 0.08 per mille in the average triple-oxygen-isotope value of shales across the Archaean–Proterozoic boundary. We suggest that our data are best explained by a shift in the nature of water–rock interactions, from near-coastal in the Archaean era to predominantly continental in the Proterozoic, accompanied by a decrease in average surface temperatures. We propose that this shift may have coincided with the onset of a modern hydrological cycle owing to the rapid emergence of continental crust with near-modern average elevation and aerial extent roughly 2.5 billion years ago.

Changes in Earth's surface environments between about 2.5 billion years ago (2.5 Gyr ago) and 2.32 Gyr ago are recorded in numerous isotopic and elemental systems, which point to a dramatic change in the oxygenation of the atmosphere and oceans at that time[9,10]. These changes were associated with a series of three to four 'Snowball Earth' glaciations[11,12], whose origin and driving forces are still debated. The major geochemical and biogeochemical rearrangements in Earth's surface environments at the Archaean–Proterozoic boundary (2.5 Gyr ago) also left numerous signatures in the geological record. Among these signatures is a steep rise in the oxygen isotopic $^{18}O/^{16}O$ ratio (expressed as $\delta^{18}O$, the deviation in the ratio in per mille relative to the standard ratio in modern seawater, VSMOW) of shales and zircons in the Late Archaean, followed by a progressively decelerating increase in these values into the Phanerozoic[13,14]. This first-order trend was modulated by the assembly and break-up of supercontinents[14]. However, it is unclear how these changes in $\delta^{18}O$ (as well as other parameters) relate to the isotopic evolution of continental crust, to the evolution of meteoric water, or to weathering conditions at Earth's surface.

We present here triple-oxygen-isotope measurements of shales, which are the dominant sedimentary rocks on Earth and the products of the chemical and physical weathering of landmasses that are exposed to the atmosphere. Shales consist mainly of clay minerals, secondary quartz and unmodified detrital minerals; studies of shales have been used previously to constrain the chemical evolution of Earth's crust through time[1,3,8,15,16]. The triple-oxygen-isotope composition of shales is expressed here as $\delta'^{18}O$ and $\Delta'^{17}O$ values; the latter parameter reflects linearized deviations in per mille (‰) of $^{17}O/^{16}O$ ratios relative to a mass-dependent $^{17}O/^{16}O$ versus $^{18}O/^{16}O$ fractionation line with a reference slope of 0.5305 (see Fig. 1 and Supplementary Information section 'Methods' for details). Both parameters are independent functions of temperature and oxygen-isotope variations and fractionation processes in surface environments[17–19].

The $\delta'^{18}O$–$\Delta'^{17}O$ signature of bulk shales is defined by: first, the proportions of detrital versus authigenic mineral components; second, the temperature of weathering and diagenesis, which affects isotopic fractionation factors; and third, the isotopic composition of the altering water (Fig. 1). Although used extensively in the past, $^{18}O/^{16}O$ ratios alone[14,20,21] are insufficient for disentangling the impact of these various processes on the shale composition. However, as we show here, the combined use of $\delta'^{18}O$ and $\Delta'^{17}O$ removes the ambiguities associated with using $\delta'^{18}O$ alone, and allows us to reconstruct past surface conditions and the composition of meteoric waters involved in weathering.

The shale samples used here are the same as those used previously[14], with the addition of 30 composite samples (formation-averaged) and 10 individual recent and Archaean samples (Fig. 1, Extended Data Tables 1–3). These 278 samples were collected from outcrops and drill holes on all continents and span 3.7 Gyr. The measured $\delta^{18}O$ values agree with previously determined values for shales and other sediment types and detrital zircons (Fig. 2)[7,13,14,16]. The bulk shales cover a large field in the $\delta'^{18}O$–$\Delta'^{17}O$ space (Fig. 1). The observed $\Delta'^{17}O$ values range from those typical of mantle and crust (−0.05‰ to −0.09‰)[18,22] to −0.3‰. Furthermore, the triple-oxygen-isotope data of shales fall on different mass-dependent fractionation lines, with slopes between the $^{18}O/^{16}O$ and $^{17}O/^{16}O$ ratios ranging from 0.529 to 0.516 (Fig. 1, Supplementary Information)—values that encompass the entire range of slopes described previously for mass-dependent processes on Earth[22].

Our results confirm a gradual trend of increasing $\delta'^{18}O$ values from 3.7 Gyr ago towards modern times. Meanwhile, the $\Delta'^{17}O$ values of shales exhibit a stepwise shift to more negative and more diverse values during and after the Archaean–Proterozoic transition (Fig. 2a). Shales older than 2.5 Gyr define an average $\Delta'^{17}O$ value of $-0.047‰ \pm 0.012‰$ ($\pm 2\sigma$), while shales younger than 2.2 Gyr—deposited in later stages and after the Great Oxidation Event (GOE) roughly 2.32 Gyr ago, during which O_2 appeared in the atmosphere—yield an average $\Delta'^{17}O$ value of $-0.118‰ \pm 0.024‰$. This difference in the triple-oxygen-isotope composition of the two age groups cannot be explained solely by different equilibration temperatures, or by the mixing of different proportions of variably weathered detrital materials, as the shale record cuts across the $\delta'^{18}O$–$\Delta'^{17}O$ trends defined by these processes (Fig. 1), requiring different initial $\delta'^{18}O$ meteoric waters. Moreover, there is no difference in the chemical index of alteration (CIA[23]) or in the proportions of minerals (determined by X-ray diffraction, XRD) of the studied shales across the GOE (Extended Data Table 2, Fig. 2) that could explain the observed shift in the $\Delta'^{17}O$ values of the shales. The invariable titanium-isotope values[8] and constant characteristic elemental ratios of the studied shales[14] (Extended Data Fig. 1) suggest that

[1]Department of Earth Sciences, University of Oregon, Eugene, OR, USA. [2]Department of Earth Sciences, University of Geneva, Geneva, Switzerland. [3]Origins Laboratory, Department of the Geophysical Sciences and Enrico Fermi Institute, The University of Chicago, Chicago, IL, USA. [4]Department of Geology, University of Johannesburg, Auckland Park, South Africa. [5]Department of Geology, Pomona College, Claremont, CA, USA. [6]Department of Earth Sciences, University of California, Riverside, CA, USA. *e-mail: bindeman@uoregon.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

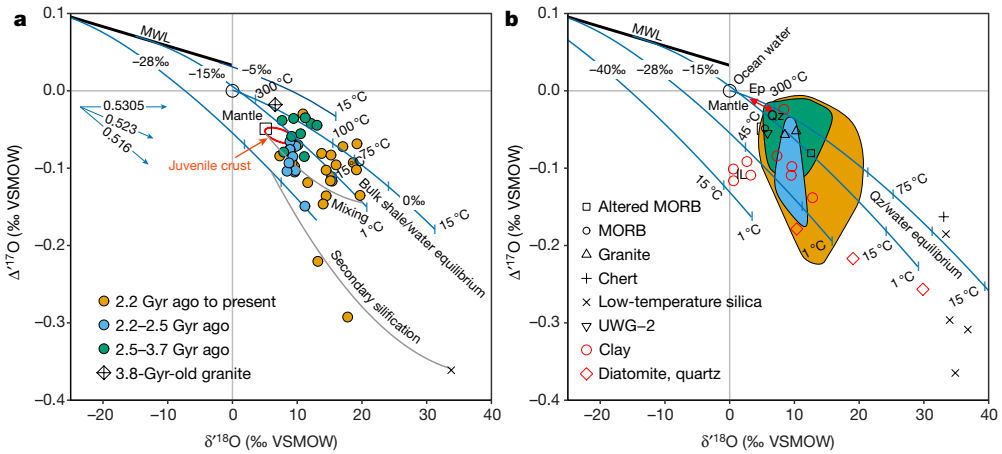


Fig. 1 | Triple-oxygen-isotope systematics of ancient and modern terrestrial materials. a, Ancient materials. **b**, Modern materials. The coloured fields in **b** correspond to the ancient-shale data shown with coloured dots in **a**. The concave blue curves represent isotopic fractionation between weathering products and meteoric water. The labels on the blue curves show the $\delta'^{18}O$ values of meteoric waters, ranging from −5‰ to −40‰, and of modern ocean water (black open circle; 0‰). The modern meteoric water line is labelled MWL. In **a**, the fractionation curves for bulk shale/water equilibration were constructed assuming a weathering product of 70% illite and 30% quartz (see Supplementary Information). In **b**, the fractionation curves are based on experimentally determined quartz/water triple-oxygen-isotope fractionations[19]. In **a**, the grey convex mixing curve connects unweathered terrestrial materials (such as mantle or upper crust) with the weathering products located on the blue fractionation lines; the grey secondary silicification curve connects modern high-$\delta'^{18}O$ and low-$\Delta'^{17}O$ materials (such as cherts and sponge spicules; individual data are shown in **b**). In **b**, the various Earth materials analysed here (red symbols) and in refs[18,19] (black symbols) are normalized to a mantle $\Delta'^{17}O$ value of −0.05‰; the double-headed red arrow, with a slope of 0.528, connects coexisting hydrothermal quartz (Qz) and epidote (Ep) from the modern deep-ocean core 504B, which formed at about 300 °C in equilibrium with roughly 0‰ seawater. In **a**, the blue arrows and associated values depict the slopes of triple-oxygen fractionation in linearized $\delta'^{17}O$–$\delta'^{18}O$ space, with the values reflecting logarithmic linearization of the delta scales[32] (see Supplementary Information); 0.5305 is characteristic of infinitely high temperatures and smaller slopes are characteristic of lower-temperature fractionations. IL, illite; VSMOW, Vienna Standard Mean Ocean Water.

the parental, atmospherically exposed continental crust undergoing weathering was similar in chemical composition to the modern crust, and has had similar proportions of mafic and felsic rocks since at least 3.5 Gyr ago, with a greater contribution of komatiites in the Archaean[8,16]. Higher-temperature Archaean oceans[24], or a greater contribution of hydrothermal clays to Archaean shales, would result in less-positive $\delta'^{18}O$ and less-negative $\Delta'^{17}O$ values. Although this could help to explain the lower $\delta'^{18}O$ values that we observe[24], it cannot explain the vertically extending trend of lower $\Delta'^{17}O$ values (Fig. 1, Extended Data Fig. 2) or the step change in $\Delta'^{17}O$ values 2.5 Gyr ago (Fig. 2). In addition, our shales show no geological or mineralogical evidence for a substantial change in hydrothermal contribution or weathering intensity in the CIA parameter across the Archaean–Proterozoic boundary[14]. Taking a cue from the modern world, where meteoric water shows variable $\delta'^{18}O$–$\Delta'^{17}O$ compositions[17], the simplest explanation for some of the oxygen-isotope variations measured in ancient shales is that they were in part inherited from the waters involved in rock alteration on the continents.

We applied recently established isotope-fractionation factors for $^{18}O/^{16}O$ and $^{17}O/^{16}O$ between quartz and water at different temperatures[19] to transform our measured, raw $\delta'^{18}O$–$\Delta'^{17}O$ data for shales into actual surface weathering conditions. We also calculated the equilibrium fractionation of oxygen isotopes between bulk shale and water, at different temperatures and initial $\delta'^{18}O$–$\Delta'^{17}O$ water values along the meteoric water line (MWL; Fig. 1, Extended Data Figs. 3–5). Oxygen-isotope fractionation between clay minerals and water under low temperatures is less than that for water and secondary quartz[20], but the two mineral/water pairs follow the same fractionation law (ref. [18]; see also Supplementary Information and Extended Data Fig. 6). The bulk shale/water fractionation lines (blue curves in Fig. 1a) define a negative slope in $\delta'^{18}O$–$\Delta'^{17}O$ space. Mixing detrital material derived from the continental crust with authigenic minerals follows a subparallel curve (grey line in Fig. 1a) within the field defined by the isotope-fractionation curves. The calculation shows that variations in temperature, in the initial oxygen-isotope composition of altering waters, and in mixing ratios between detrital material and weathering products can explain the overall trend and negative co-variations in our data.

The proportion of weathering products in a shale can be assessed independently via its mineralogical composition (through XRD) and/or its chemical composition (such as through the CIA[23]). Combining this estimate with the $\delta'^{18}O$ and $\Delta'^{17}O$ values of the shales and the pristine detrital components (igneous rocks) allows us to calculate, by mass balance, the $\delta'^{18}O$ and $\Delta'^{17}O$ values of the weathering products (Extended Data Fig. 3). The CIA index has remained nearly constant through time[14] (Extended Data Fig. 7), suggesting that a secular trend in weathering intensity is unlikely to introduce a systematic bias in this approach. The equations for isotopic fractionation during water–rock interactions and the equation of the MWL in $\delta'^{18}O$–$\Delta'^{17}O$ space (Fig. 1) can be used to independently calculate the water–rock interaction temperature and the oxygen-isotopic composition ($\delta'^{18}O_W$ and $\Delta'^{17}O_W$) of waters involved in weathering and diagenesis (Extended Data Figs. 3, 4). This approach is likely to be oversimplified, because detrital components in shale precursors were probably altered by a range of meteoric and diagenetic waters at different temperatures in watersheds. But, given that we compare shales with shales, all of which have comparable CIAs, these complexities do not affect the first-order interpretations afforded by quantitative modelling.

We find that the temperatures of interacting water and rock derived from the inversion of shale $\delta'^{18}O$–$\Delta'^{17}O$ values were higher during the Archaean than after it (Extended Data Fig. 5). This modelling exercise also shows that—although our measurements reveal trends towards heavier average $\delta'^{18}O$ and lighter average $\Delta'^{17}O$ values in shales over time—the waters involved in surface alteration and weathering processes became lighter in $\delta'^{18}O$, heavier in $\Delta'^{17}O$, and more variable in both $\delta'^{18}O$ and $\Delta'^{17}O$ after roughly 2.5 Gyr ago. Although the quantitative analysis makes important simplifications, it does capture the essential features of weathering conditions on continents through time.

Another explanation for the stepwise change in $\Delta'^{17}O$ values during the GOE could be the appearance of atmospheric oxygen and ozone.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

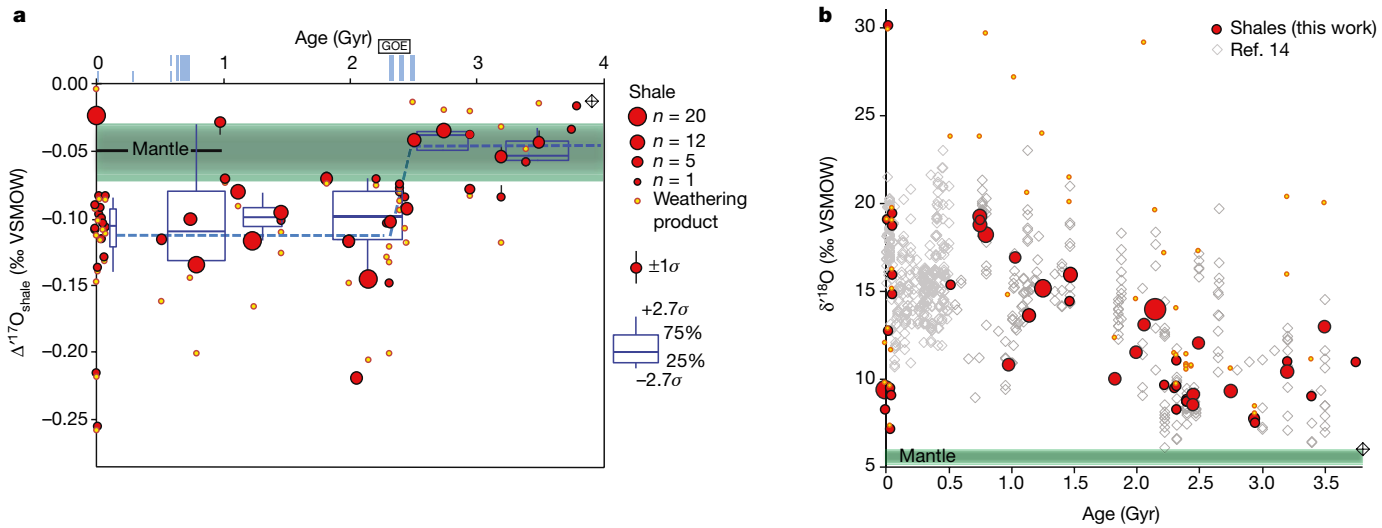


Fig. 2 | Oxygen-isotopic compositions of shales through time. a, $\Delta'^{17}O$ record of shales (red points) and calculated oxygen-isotopic composition of weathering products (yellow points), showing a stepwise change in composition between 2.5 Gyr ago and 2.2 Gyr ago. The blue dashed lines represent the mean $\Delta'^{17}O$ values, weighted by the composite sample size (n), for shales before 2.5 Gyr ago and after 2.2 Gyr ago. The boxes represent the medians, interquartile ranges and extreme values (see legend). A t-test reveals that the Archaean and post-Archaean $\Delta'^{17}O$ values are statistically distinct ($0.004 < P < 0.02$, well below a statistical significance value of 0.05). We attribute the different $\delta'^{18}O$ and $\Delta'^{17}O$ values of pre- and post-Archaean shales in this diagram to a change from a coast-dominated to a more continental hydrological cycle and weathering conditions (see text for details). The blue vertical bars at the top indicate major glacial episodes. **b**, $\delta'^{18}O$ record of shales (red points) from our dataset superimposed on the dataset from ref. [14] (white diamonds).

The formation of ozone with highly positive $\Delta'^{17}O$ values (+30‰ to +100‰) via ultraviolet photolysis in the stratosphere leaves atmospheric oxygen with the slightly negative $\Delta'^{17}O$ value of −0.3‰ in today's 21 vol% oxygen atmosphere[22,25], but probably much less in Proterozoic conditions of less than 1 vol% oxygen. However, not only was the atmospheric $\Delta'^{17}O$ signal small, there is also no mechanism by which to transfer this signature into the meteoric water cycle and hence into crustal silicate weathering products such as shales.

As argued above, we favour an explanation by which the difference in the $\Delta'^{17}O$ values of pre- and post-GOE shales occurred through a change in the meteoric water cycle (Fig. 3). Starting roughly at the Archaean–Proterozoic transition, the emerged crust would have interacted with waters that had more variable and on average more negative $\delta'^{18}O_W$ values and more positive $\Delta'^{17}O_W$ values than before the GOE (with $\Delta'^{17}O_W$ shifted by approximately +0.1‰). The observed shift in the triple-oxygen-isotope composition of shales would also have required lower post-GOE surface temperatures (Extended Data Fig. 5). This is broadly consistent with the findings of previous studies of $\delta'^{18}O$ values in cherts[24].

In the modern world, the oxygen-isotope composition of precipitation depends on the cumulative history of water loss from the air parcel that is travelling inland away from the coasts towards higher latitudes and higher altitudes (http://www.waterisotopes.org), resulting in lower $\delta'^{18}O_W$ values, higher $\Delta'^{17}O_W$ values, and more diverse compositions overall[17] (as in the MWL on Fig. 1). This combined effect, which we call 'continentality' (Fig. 3), is shown through the shale record's step change that coincides with the Archaean–Proterozoic boundary (Figs. 2a, 3). It is most likely that the observed change in the shale triple-oxygen-isotope record reflects the appearance of larger continents (Fig. 3) and higher elevations from the Proterozoic onwards—a period that is broadly contemporaneous with the final stages in the assembly of the first documented supercontinent, Kenorland[2,26], or with the formation of several supercratons[12,15] immediately before the GOE[27]. Supercontinent assembly and orogenic events result in high mountain ranges and plateaus, as occurred when India collided with Asia, forming the Himalayas and Tibet. Rising mountains in even quite low to mid latitudes result in precipitation with very light $\delta'^{18}O$ values (http://www.waterisotopes.org) that correlate with elevation, with a roughly 2‰–3‰ drop in $\delta'^{18}O$ per kilometre of altitude gain[28].

Supporting our interpretations of triple-oxygen variations in shales is the strontium-isotope record of marine carbonates, which suggests

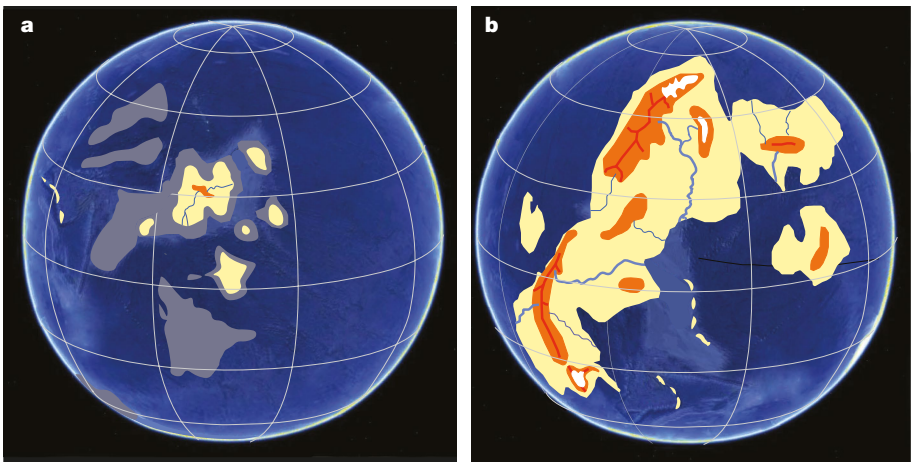


Fig. 3 | Conceptual palaeohypsometry of Archean and Proterozoic worlds. These images are based on palaeomagnetic and tectonic reconstructions (see refs [12,26] and references therein). **a**, The late Archaean era (shortly before 2.5 Gyr ago) during the assembly of the supercontinent Kenorland. **b**, The early/mid Paleoproterozoic era, after the occurrence of the GOE. The oceans are shallower in **a** compared with **b**, and the (excessively flooded) continents are smaller and lower, resulting in different hydrologic and weathering cycles, as described here.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

that the area of emerged continental crust increased substantially and irreversibly at roughly the Archaean–Proterozoic boundary[29]. From a geodynamic perspective, models of a cooling Earth call for a thickening of the lithosphere and the establishment of a higher continental freeboard by about 2.5 Gyr ago owing to increased mantle viscosity[5,30,31]. The emergence of large landmasses (Fig. 3) would also have led to a larger weathering sink for carbon dioxide, which occurred at greater concentrations in the Archaean, resulting in a transition to moderate surface temperatures after the GOE. We conclude that this set of large-scale tectonic and near-surface changes best explains the observed shift in the $\delta'^{18}O$–$\Delta'^{17}O$ composition of shales between 2.5 Gyr ago and 2.2 Gyr ago (Fig. 2).

The rapid increase in Earth's subaerial surface and overall hypsometry 2.5 Gyr ago that we infer here (Fig. 3) would have also increased Earth's albedo, the flux of nutrients to the oceans[8] from continents undergoing subaerial weathering, and the extent of continental margins, additionally resulting in a higher rate of burial of organic carbon and reducing the concentration of carbon dioxide in the air. Together, these changes could have contributed to the cooling of the planet and to the snowball glaciations of the early Palaeoproterozoic, followed by the GOE, highlighting how Earth's interior could have influenced surface redox conditions and chemistry. The most dramatic change in Earth's history was marked by a transition from hot and largely anoxic surface conditions to an oxygenated atmosphere with moderate surface temperatures. Our study suggests that this transition might have been modulated by long-term cooling of the subcontinental mantle and lithosphere, rendering it capable of supporting a thicker crust[1,2,30,31]. This would have led to the emergence of extensive landmasses at the Archaean–Proterozoic boundary, with life and surface conditions adjusting to—rather than triggering—the change in atmospheric oxygen concentrations.

Data availability

Data are provided in Extended Data Tables 1–3.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0131-1.

Received: 2 October 2017; Accepted: 7 March 2018;
Published online 23 May 2018.

1. Taylor, S. R. & McLennan, S. M. The geochemical evolution of the continental crust. *Rev. Geophys.* **33**, 241–265 (1995).
2. Condie, K. C. *Plate Tectonics and Crustal Evolution* 3rd edn (Pergamon Press, Oxford, 2013).
3. Belousova, E. A. et al. The growth of the continental crust: constraints from zircon Hf-isotope data. *Lithos* **119**, 457–466 (2010).
4. Keller, C. B., Schoene, B., Barboni, M., Samperton, K. M. & Husson, J. M. Volcanic–plutonic parity and the differentiation of the continental crust. *Nature* **523**, 301–307 (2015).
5. Dhuime, B., Wuestefeld, B. & Hawkesworth, C. J. Emergence of modern continental crust about 3 billion years ago. *Nat. Geosci.* **8**, 552–555 (2015).
6. Lee, C.-T. A. et al. Deep mantle roots and continental emergence: implications for whole-Earth elemental cycling, long-term climate, and the Cambrian explosion. *Int. Geol. Rev.* **60**, 431–448 (2017).
7. Hawkesworth, C. J., Cawood, P. A., Dhuime, B. & Kemp, A. I. S. Earth's continental lithosphere through time. *Annu. Rev. Earth Planet. Sci.* **45**, 169–198 (2017).
8. Greber, N. D. et al. Titanium isotopic evidence for felsic crust and plate tectonics 3.5 billion years ago. *Science* **357**, 1271–1274 (2017).
9. Farquhar, J. & Wing, B. A. The terrestrial record of stable sulphur isotopes: a review of the implications for evolution of Earth's sulphur cycle. *Geol. Soc. Lond. Spec. Publ.* **248**, 167–177 (2005).
10. Bekker, A. et al. Dating the rise of atmospheric oxygen. *Nature* **427**, 117–120 (2004).
11. Hoffman, P. F. The Great Oxidation and a Siderian snowball Earth: MIFS based correlation of Paleoproterozoic glacial epochs. *Chem. Geol.* **362**, 143–156 (2013).
12. Gumsley, A. P. et al. Timing and tempo of the Great Oxidation Event. *Proc. Natl Acad. Sci. USA* **114**, 1811–1816 (2017).
13. Valley, J. W. et al. 4.4 billion years of crustal maturation: oxygen isotope ratios of magmatic zircon. *Contrib. Mineral. Petrol.* **150**, 561–580 (2005).
14. Bindeman, I. N., Bekker, A. & Zakharov, D. O. Oxygen isotope perspective on crustal evolution on early Earth: a record of Precambrian shales with emphasis on Paleoproterozoic glaciations and Great Oxygenation Event. *Earth Planet. Sci. Lett.* **437**, 101–113 (2016).
15. Bleeker, W. The late Archean record: a puzzle in ca. 35 pieces. *Lithos* **71**, 99–134 (2003).
16. Gaschnig, R. M. et al. Compositional evolution of the upper continental crust through time, as constrained by ancient glacial diamictites. *Geochim. Cosmochim. Acta* **186**, 316–343 (2016).
17. Luz, B. & Barkan, E. Variations of $^{17}O/^{16}O$ and $^{18}O/^{16}O$ in meteoric waters. *Geochim. Cosmochim. Acta* **74**, 6276–6286 (2010).
18. Pack, A. & Herwartz, D. The triple oxygen isotope composition of the Earth mantle and understanding $\delta^{17}O$ variations in terrestrial rocks and minerals. *Earth Planet. Sci. Lett.* **390**, 138–145 (2014).
19. Sharp, Z. D. et al. A calibration of the triple oxygen isotope fractionation in the SiO_2-H_2O system and applications to natural samples. *Geochim. Cosmochim. Acta* **186**, 105–119 (2016).
20. Savin, S. & Epstein, S. The oxygen and hydrogen isotope geochemistry of ocean sediments and shales. *Geochim. Cosmochim. Acta* **34**, 43–63 (1970).
21. Land, L. S. & Lynch, F. L. $\delta^{18}O$ values of mudrocks: more evidence for an ^{18}O-buffered ocean. *Geochim. Cosmochim. Acta* **60**, 3347–3352 (1996).
22. Bao, H. M., Cao, X. B. & Hayles, J. Triple-oxygen isotopes: fundamental relationships and applications. *Annu. Rev. Earth Planet. Sci.* **44**, 463–492 (2016).
23. Nesbitt, H. W. & Young, G. N. Early Proterozoic climates and plate motions inferred from major element chemistry of lutites. *Nature* **299**, 715–717 (1982).
24. Knauth, L. P. & Lowe, D. R. High Archean climatic temperature inferred from oxygen isotope geochemistry of cherts in the 3.5 Ga Swaziland Supergroup, South Africa. *Geol. Soc. Bull.* **115**, 566–580 (2003).
25. Young, E., Yeung, L. Y. & Kohl, I. On the ^{17}O budget of atmosphere. *Geochim. Cosmochim. Acta* **135**, 102–125 (2014).
26. Mertanen, S. & Pesonen, L. J. in *From the Earth's Core to Outer Space* (ed. Haapala, I.) 11–35 (Springer, Berlin, 2012).
27. Barley, M. E., Bekker, A. & Krapez, B. Late Archean to early Paleoproterozoic global tectonics, environmental change and the rise of atmospheric oxygen. *Earth Planet. Sci. Lett.* **238**, 156–171 (2005).
28. Rowley, D. B., Pierrehumbert, R. & Currie, B. A new approach to stable isotope-based paleoaltimetry: implications for paleoaltimetry and paleohypsometry of the High Himalaya since the Late Miocene. *Earth Planet. Sci. Lett.* **188**, 253–268 (2001).
29. Flament, N., Coltice, N. & Rey, P. F. A case for late-Archaean continental emergence from thermal evolution models and hypsometry. *Earth Planet. Sci. Lett.* **275**, 326–336 (2008).
30. Vlaar, N. J. Continental emergence and growth on a cooling earth. *Tectonophysics* **322**, 191–202 (2000).
31. Korenaga, J., Planavsky, N. J. & Evans, D. A. D. Global water cycle and the coevolution of the Earth's interior and surface environment. *Phil. Trans. R. Soc. A* **375**, 20150393 (2017).
32. Miller, M. F. Isotopic fractionation and the quantification of ^{17}O anomalies in the oxygen three-isotope system: an appraisal and geochemical significance. *Geochim. Cosmochim. Acta* **66**, 1881–1889 (2002).

Acknowledgements This work was supported by National Science Foundation (NSF) grant EAR1447337 and by the University of Oregon. N.D. was supported by NSF grant EAR1502591. A.B. was supported by National Sciences and Engineering Research Council (NSERC) Discovery and Accelerator grants. We thank P. Hoffman, S. Mertanen and D. Evans for discussions about Precambrian palaeogeography and environmental changes; K. Johnson for technical help with vacuum lines; and O. Melnik for help with programming.

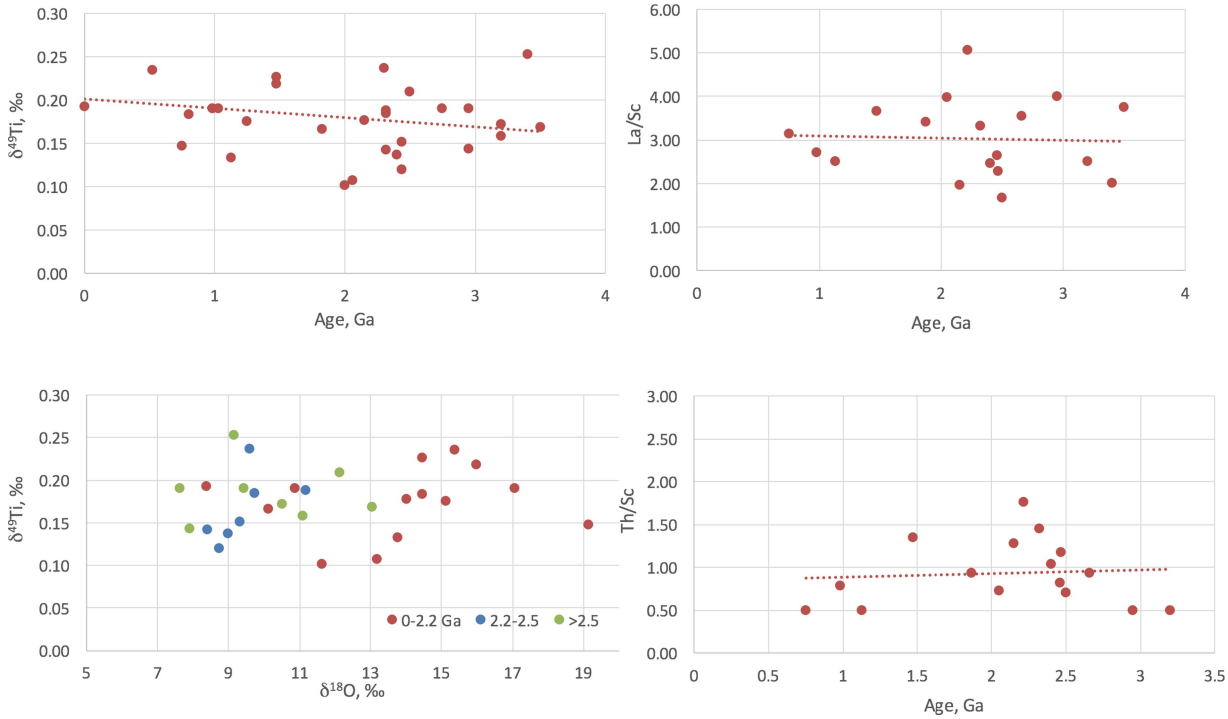
Reviewer information *Nature* thanks C. Hawkesworth and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions I.N.B. conceived the study and wrote the paper; D.O.Z., J.P. and I.N.B. built the purification line and collected the data; N.D.G. and N.D. provided composite shale samples studied previously for titanium isotopes; A.B., A.H. and G.J.R. provided samples; J.S.L. performed major- and trace-element analyses; I.N.B. and N.D. discussed the inversion approach and I.N.B. implemented the model; D.O.Z. contributed to statistical treatment; A.B., N.D., N.D.G. and G.J.R. contributed to discussions on Precambrian environments and crustal evolution. All authors contributed to writing and editing the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0131-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0131-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to I.N.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

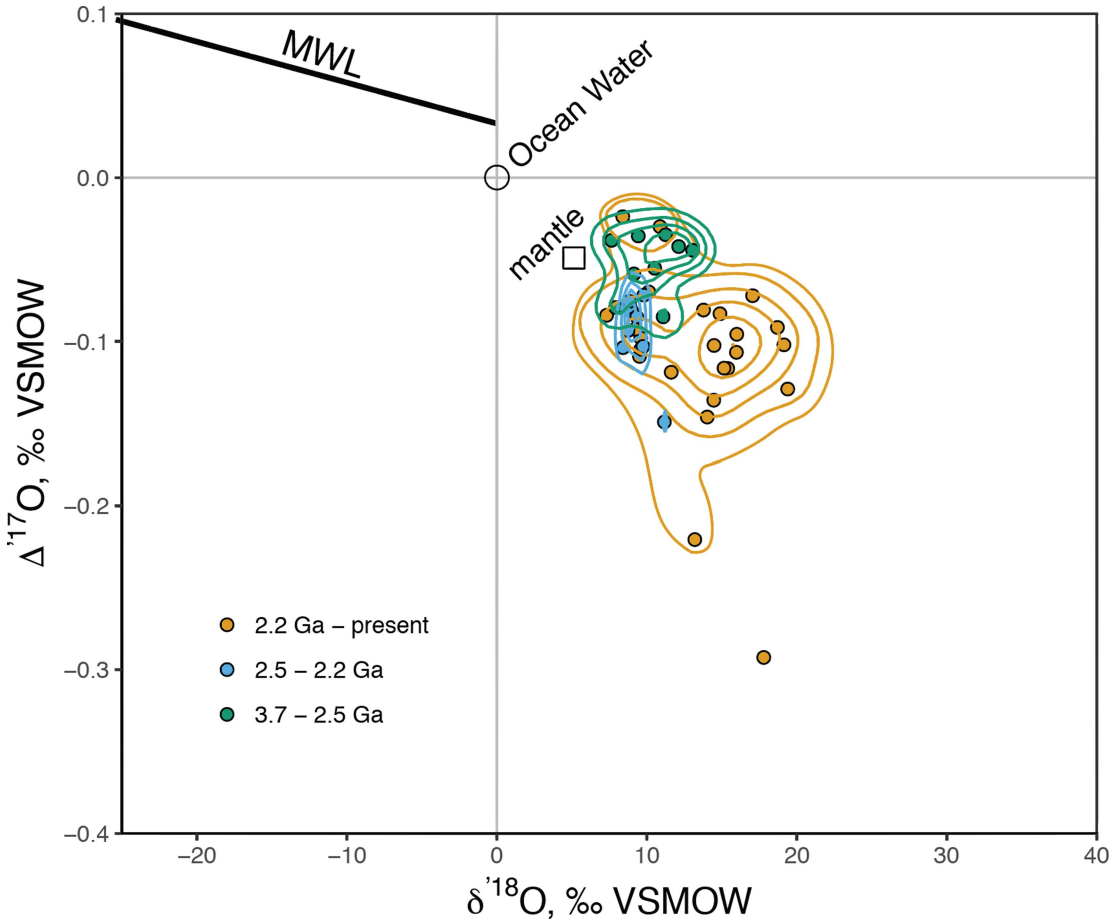
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 1 | Comparison of isotopic and key elemental ratios of the shales studied here, illustrating the relative constancy of the composition of the exposed crust that is undergoing weathering, and in particular proportion of exposed mafic versus silicic rocks. See, for example, refs [1,33]. The δ^{49}Ti data are from ref. [8], which used a large dataset that included many of the samples studied here. The elemental data are from Extended Data Table 3 and ref. [14]. Variation in the composition of the exposed crust cannot explain the oxygen-isotope trends that we have identified here.

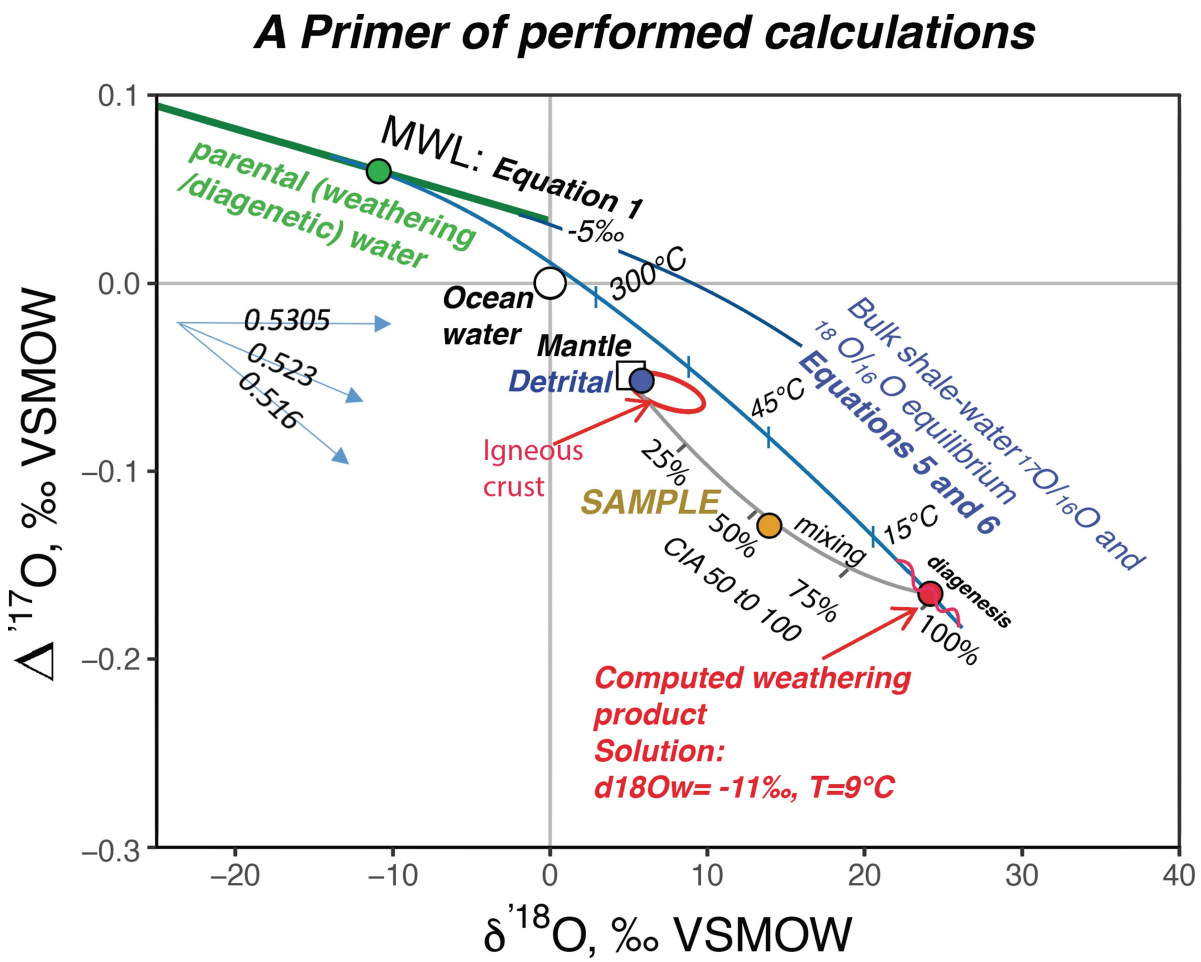
33. McLennan, S. M., Hemming, S., McDaniel, D. K. & Hanson, G. N. Geochemical approaches to sedimentation, provenance and tectonics. *Geol. Soc. Am. Spec. Pap.* **284**, 21–40 (1993).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



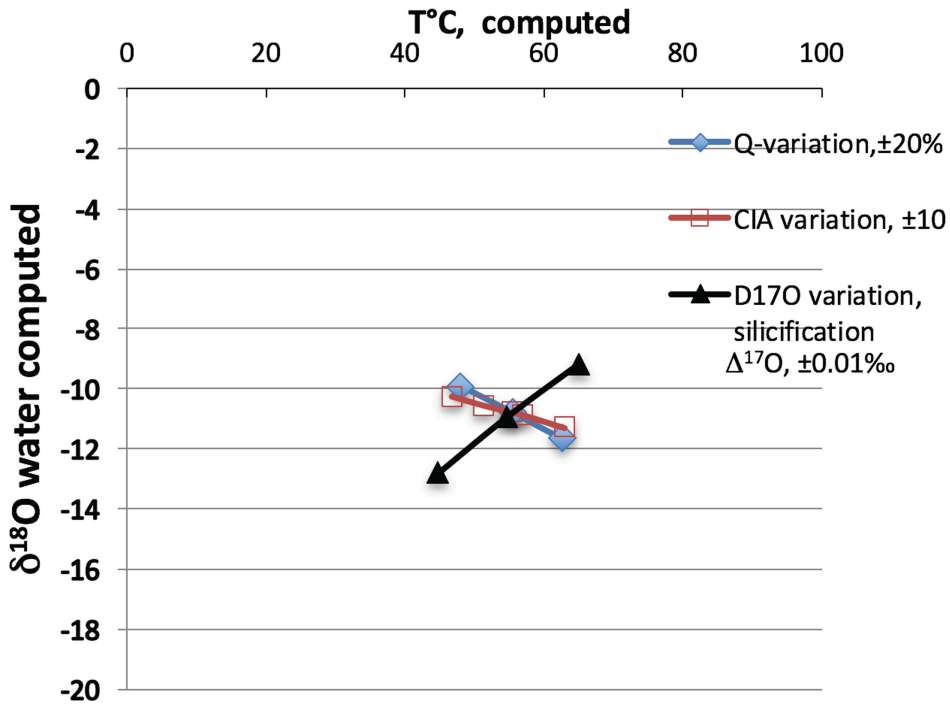
Extended Data Fig. 2 | Distribution density plot for the samples studied here. Data from Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



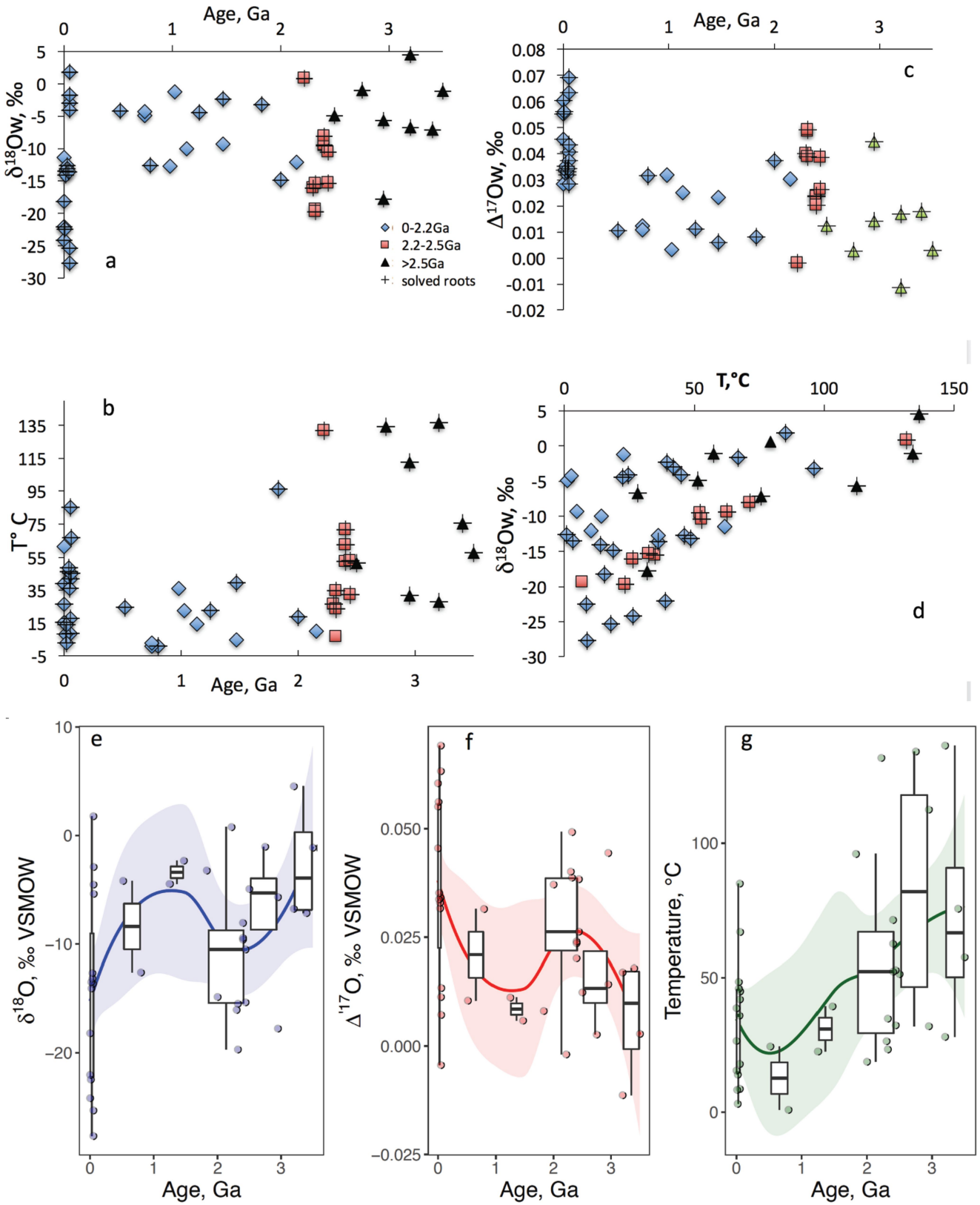
Extended Data Fig. 3 | Calculating the $\delta'^{18}O$ and $\Delta'^{17}O$ values of the weathering products of shales. Solving a system of three unknowns (for example, the temperature of alteration and the values of $\Delta'^{17}O$ and $\delta'^{18}O$ along the MWL) in three equations (Supplementary Information equations (1), (5) and (6)) follows these steps: the results of Supplementary Information equations (1), (5) and (6) are substituted sequentially into each other until one unknown is left. This results in a function with respect to temperature (T, a single parameter) to solve: $y = 0 = 36.323T - 23.33T^2 + 0.00264T^3 - 1.38 \times 10^7$. This is a third-order polynomial equation that has three roots; we solve for roots in a realistic temperature range to obtain the temperature and $\delta^{18}O_W$. The concave blue curve originating from a point on the MWL shows isotope fractionation between shale and water with the indicated temperatures of equilibration. The grey curve represents a mixing line between detrital and weathering products of the indicated proportions, computed using the CIA.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 4 | Evaluating the sensitivity of reconstructed $\delta^{18}O_W$ and temperature values to variations in the input parameters. We show here the effects of varying the proportion of quartz (Q) in shales (±20%), the CIA (±10 units), and $\Delta'^{17}O$ (D170) (±0.01‰), the latter occurring naturally owing to, for instance, silicification.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



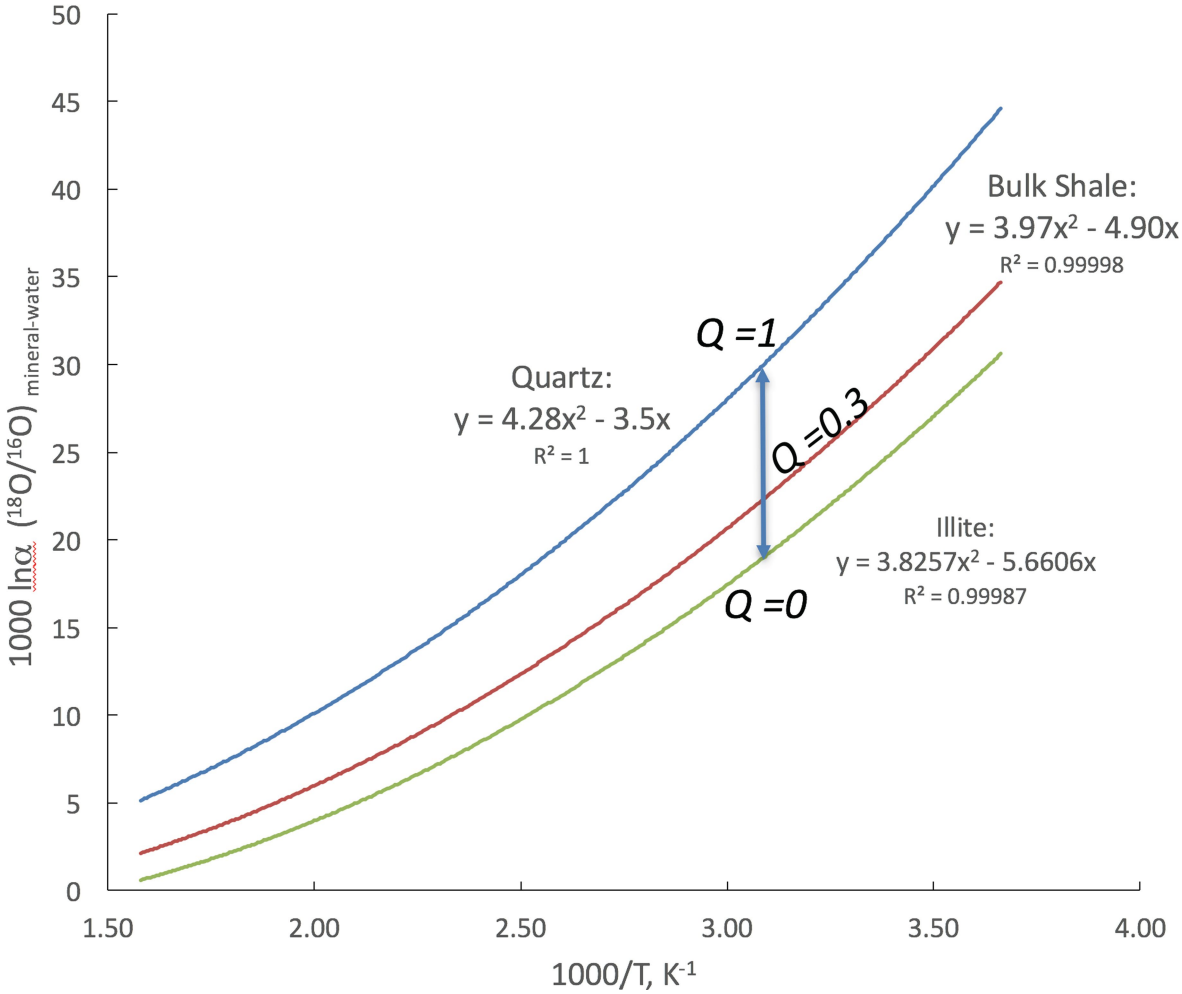
Extended Data Fig. 5 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 5 | Calculated temperature, $\delta^{18}O_W$ and $\Delta^{17}O_W$ values plotted against age of shales. a, $\delta^{18}O_W$ versus age; **b**, temperature versus age; **c**, $\Delta^{17}O_W$ versus age; **d**, $\delta^{18}O_W$ versus temperature. These values are based on the solution of three equations with three input parameters: $\delta'^{18}O_{shale}$, $\delta'^{17}O_{shale}$ and the proportion of quartz (see Extended Data Fig. 3). Sensitivity analysis is provided in Extended Data Fig. 4. Most measurements yielded solved roots (plus signs within symbols); when equations could not be solved, small variations in the input parameters (Extended Data Table 1) allowed us to find roots. In particular, correcting for secondary silicification (see Extended Data Table 1) by decreasing $\Delta'^{17}O_{shale}$ by 0.01‰ to 0.08‰ along the silicification line allowed us to find roots in all but two cases. Note that the overall calculated $\delta^{18}O_W$ and temperature ranges agree with modern and recent values for surface, diagenetic, basinal and pore waters measured in drillholes[21]. Note also that the absolute calculated values for the temperature of water–rock interaction (weathering) and $\delta^{18}O_W$ depend on the assumed isotopic fractionations in Extended Data Fig. 6; however, given that these values solved within realistic bounds, the fractionations of ref. [19] and the MWL defined in ref. [17] are probably well constrained and calibrated in absolute triple-oxygen-isotope space. The lowest $\delta^{18}O_W$ and temperature are computed for recent clay samples from Antarctica, and for 2.5–2.2-Gyr-old synglacial Palaeoproterozoic shales, confirming the participation of low-$\delta^{18}O_w$ synglacial waters in diagenesis, as proposed previously[14]. The highest recent temperature and $\delta^{18}O_W$ values are for Palaeocene–Eocene (55-million-year-old) thermal maximum shales (ref. [34] and Extended Data Table 1). **e**–**g**, Interquartile range statistics and running averages for the parameters computed in **a**–**d**.

34. Frieling, J. et al. Paleocene–Eocene warming and biotic response in the epicontinental West Siberian Sea. *Geology* **42**, 767–770 (2014).

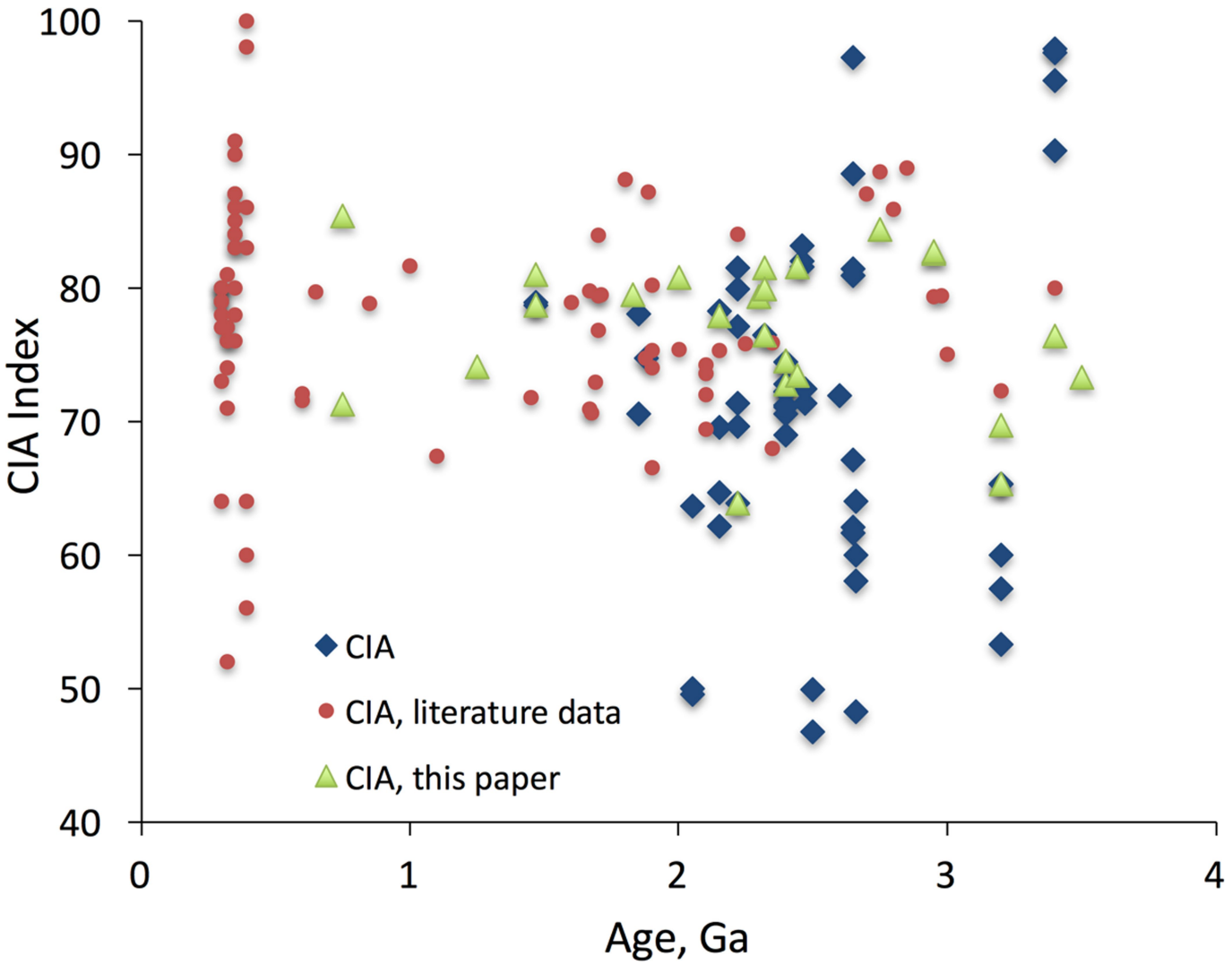
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Quartz/water, illite/water and bulk-shale/water $^{18}O/^{16}O$ fractionation factors. The quartz/water and illite/water 1,000lnα ($^{18}O/^{16}O$) fractionation factors are based on refs [19,35]. The bulk-shale/water 1,000lnα ($^{18}O/^{16}O$) fractionation factors are based on the assumption that bulk shale comprises 70% illite and 30% quartz (that is, Q = 0.3). The blue line (for quartz/water fractionation) corresponds to Supplementary Information equation (2) and equation (9) in ref. [19]. The green line (for illite/water fractionation) is the best-fit second-order polynomial with two fit coefficients based on the equation in ref. [35] that includes three fit coefficients. We used these coefficients to solve equations for bulk-shale/water triple-oxygen-isotope fractionations according to the proportion of quartz (determined through X-ray-diffraction; Extended Data Table 2).

35. Zheng, Y. F. Calculation of oxygen isotope fractionation in hydroxyl-bearing silicates. *Earth Planet. Sci. Lett.* **120**, 247–263 (1993).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | Chemical index of alteration (CIA) plotted against time. $CIA = Al_2O_3/(Al_2O_3 + CaO + Na_2O + K_2O)_{mol}$. Data are taken from this work (triangles), ref. [14] (diamonds), and literature data (circles).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Triple-oxygen-isotope analyses of the shale samples used here

	measured				linearized and normalized					Weathering product								Weathering waters					
AGE, Ga	$\delta^{17}O$, ‰, av	$\delta^{18}O$, ‰, av	$\Delta^{17}O$, ‰, av	±stdev	$\delta'^{17}O$, ‰, norm *	$\delta'^{18}O$, ‰, norm	$\Delta'^{17}O$, ‰, norm	n	CIA‡	$\delta^{18}O$ CIA corr ¶	$\Delta'^{17}O$ CIA corr	$\delta'^{17}O$ CIA corr	drillcore (d), outcrop (o)	Composite (cs), individual shale (is)	Sample size, n	Ref	Sample Description	$\delta^{65}Ti$ †	Quartz	T, °C	$\delta^{18}Ow$, ‰	$\delta'^{17}Ow$ ‰	$\Delta'^{17}Ow$ ‰
0	4.269	8.238	-0.101	0.022	4.426	8.387	-0.024	2	*75*	9.775	0.022	5.208	d	cs	18	36	Carribean Sea, Core 34 of Swedish Deep Sea Expedition 1947, secondary carbonate present	0.192	0.3	62	-11.4	-6.0	0.029
0	4.786	9.378	-0.189	0.011	4.940	9.518	-0.109	3	*75*	12.035	-0.148	6.236	o	is	1		Holocene silt, Oregon		0.3	15	-18.2	-9.6	0.046
0.00	1.155	2.480	-0.160	0.031	1.320	2.661	-0.091	3	100	2.661	-0.091	1.320	o	is	1		Suislaw River, Oregon clay +silt , river mud, Florence, Oregon		0.3	39	-22.0	-11.6	0.055
0.00	1.450	3.070	-0.177	0.002	1.615	3.249	-0.109	2	100	3.249	-0.109	1.615	o	is	1		Suislaw River, Oregon clay +silt , Florence, Oregon		0.3	27	-24.2	-12.8	0.060
0.05	0.030	0.376	-0.17	0.015	0.196	0.560	-0.101	2	100	0.560	-0.101	0.196	o	is	1	63	sample 2367, Antarctic claystone, Plio-Pleistocene, Transantarctic Mt, 85°S		0.01	18	-25.3	-13.4	0.063
0.05	0.030	0.405	-0.16	0.019	0.196	0.588	-0.116	5	100	0.588	-0.116	0.196	o	is	1	63	sample 2367, Antarctic claystone, Plio-Pleistocene, Transantarctic Mt, 85°S		0.01	9	-27.7	-14.6	0.069
0.01	9.775	19.053	-0.333	0.005	9.893	19.057	-0.217	3	*100*	19.057	-0.217	9.893	o	is	1	64	F108534_A, Late Miocene, Nevada, non-marine diatomite		0.99	8	-22.5	-11.9	0.056
0.018	15.533	30.112	-0.441	0.008	15.580	29.852	-0.257	3	*100*	29.852	-0.257	15.580	o	is	1	38	locality UO11095 Cabrillo Beach Museum parking lot, California: Valmonte Diatomite: middle Miocene		0.99	3	-13.5	-7.1	0.034
0.02	6.513	12.706	-0.228	0.012	6.657	12.809	-0.138	4	*100*	12.809	-0.138	6.657	o	is	1	39	R3888, kaolinite of lateritic profile, Long Reef Beach, near Deewhy, NSW, Australia: middle Miocene		0.01	14	-14.1	-7.4	0.035
0.035	3.634	7.150	-0.159	0.021	3.793	7.308	-0.084	3	*100*	7.308	-0.084	3.793	o	is	1		John Day fossil beds, Oregon, kaolinite. Oligocene		0.01	46	-12.6	-6.7	0.032
0.035	4.836	9.452	-0.178	0.013	4.990	9.591	-0.098	3	*100*	9.591	-0.098	4.990	o	is	1		John Day fossil beds, Oregon, smectite, Oligocene		0.3	48	-13.2	-6.9	0.033
0.045	4.647	9.084	-0.172	0.007	4.802	9.227	-0.093	3	74.0	11.640	-0.117	6.058	o	cs	7	40	Tyee Fm, Triangle Lake area, Oregon Coastal Ranges		0.3	36	-13.6	-7.2	0.034
0.051	9.739	18.709	-0.186	0.017	9.839	18.720	-0.091	4	*100*	18.720	-0.091	9.839	d	is	1	34	Russkaya Polyana Fm, claystone, W Siberia, Omsk Region, well 10, EECO excursion layer, depth 195m		0.3	85	1.8	0.9	-0.004
0.052	10.060	19.391	-0.227	0.016	10.157	19.389	-0.129	3	*100*	19.389	-0.129	10.157	d	is	1	34	Russkaya Polyana Fm, claystone, W Siberia, Omsk Region, well 10, EECO excursion layer, depth 198m		0.3	42	-4.5	-2.4	0.011
0.055	7.689	14.798	-0.161	0.027	7.804	14.873	-0.083	4	*100*	14.873	-0.083	7.807	d	is	1	34	Lulinvor Formation, W Siberia, Omsk region, well 10, PETM excursion claystone layer depth 236.85		0.3	67	-2.9	-1.5	0.007
0.056	8.314	15.930	-0.137	0.016	8.375	15.988	-0.107	3	*100*	15.988	-0.107	8.375	d	is	1	34	Lulinvor Formation, W Siberia, Omsk region, well 10, PETM excursion claystone layer depth 236.9		0.3	45	-5.4	-2.8	0.013
0.52	7.908	15.312	-0.215	0.007	8.042	15.380	-0.117	3	*75*	23.759	-0.163	12.441	o	cs	2	41	Poleta Formation, Poleta Folds, Eastern California, USA	0.235	0.3	24	-4.2	-2.2	0.010
0.75	9.022	17.767	-0.403	0.005	9.148	17.795	-0.293	3	85.4	22.247	-0.385	11.417	o	cs	12	42	Cerro Espuelitas Formation, Arroyo Del Soldado Group, Uruguay, numeric age tentative	0.318	0.3				no roots
0.8	7.400	14.385	-0.231	0.014	7.538	14.466	-0.1359	3	75.0	21.932	-0.202	11.433	d	cs	16	43,44,14	Wynniatt Fm, Shaler Supergroup, Victoria Island, N Canada, GNME drill core 07-04, sec. carbonate present	0.184	0.3	1	-12.6	-6.7	0.032
0.75	9.937	19.140	-0.217	0.024	10.053	19.143	-0.102	2	71.3	35.504	-0.145	18.690	d	cs	13	14, 44	Mwashya Subgroup, Upper Roan Group, Katanga Supergroup, Zambia, Drillhole RCB1, below glacials	0.147	0.3	1	-4.9	-2.6	0.012
0.98	5.589	10.748	-0.113	0.014	5.739	10.874	-0.030	2	*75*	14.748	0.010	7.834	o	cs	8	45,46,14	Upper part of the Ui Group, Lakhanda Series, Uchur-Maya Region, Siberia, Russia	0.190	0.3	36	-12.7	-6.7	0.032
1.025	8.852	17.021	-0.177	0.012	8.979	17.061	-0.072	2	*75*	27.122	-0.074	14.314	o	cs	5	45,46,14	Neryuen Formation, Lakhanda Series, Uchur-Maya Region, Siberia, Russia	0.190	0.3	23	-1.2	-0.6	0.003
1.13	7.086	13.684	-0.173	0.009	7.227	13.775	-0.081	2	*75*	20.550	-0.092	10.810	o	cs	12	14, 37	Bylot Supergroup, northern Baffin and Bylot Islands, Canada, secondary carbonate present	0.133	0.3	14	-10.0	-5.3	0.025
1.25	7.774	15.058	-0.214	0.020	7.910	15.129	-0.116	8	74.0	23.936	-0.167	12.532	o	cs	18	47,45,14	Totta, Talyn, Trekhgornaya and Dim formations, Uchur-Maya Region, Yudoma-Maya trough, Siberia, Russia	0.175	0.3	23	-4.5	-2.4	0.011
1.47	8.260	15.940	-0.197	0.022	8.391	15.998	-0.096	2	81.2	21.420	-0.111	11.252	d	cs	14	14	Newland Formation, Lower Belt Supergroup, Montana, USA, drill core SC-93	0.218	0.11	5	-9.3	-4.9	0.023
1.47	7.441	14.398	-0.198	0.0048	7.579	14.479	-0.102	4	78.7	20.039	-0.127	10.504	d	is	1	14	SC93-1906, Newland Fm, Lower Belt Supergroup, Montana, drill core SC-93, depth 1906, studied by XRD	0.226	0.11	39	-2.3	-1.2	0.006
1.83	5.148	9.988	-0.151	0.021	5.300	10.122	-0.070	2	79.5	12.292	-0.070	6.451	d	cs	11	48, 14	Rove Formation, near Thunder Bay, Ontario, Canada, drill core 89-MC-1	0.166	0.3	96	-3.2	-1.7	0.008
2.0	5.905	11.517	-0.205	0.012	6.053	11.634	-0.119	2	80.8	14.523	-0.149	7.556	d	cs	10	49, 14	Maraloou Formation, southeastern Capricorn Orogen, Australia, drill core KDD1	0.101	0.3	19	-14.8	-7.8	0.037
2.06	6.636	13.098	-0.312	0.009	6.780	13.196	-0.221	4	64.0	29.129	-0.609	14.844	d	cs	14	51	Zaonega Formation, Lower Ludikovian Series, Karelia, Russia, drill cores 175 and 5190	0.107	0.3				no roots
2.15	7.152	13.934	-0.240	0.006	7.292	14.022	-0.146	3	77.9	19.583	-0.206	10.182	d	cs	20	51, 14	Sengoma Argillite Formation, Pretoria Group, Lobatse, Botswana, drill core STRAT 2	0.177	0.3	10	-12.1	-6.4	0.030
2.30	4.830	9.453	-0.185	0.018	4.984	9.592	-0.105	4	79.3	11.424	-0.130	5.931	d	cs	10	51,52,14	Timeball Hill Fm, Lower Pretoria Group, Transvaal Basin, S Africa, drill cores EBA-1 and 2, above all glacials	0.237	0.3	26	-16.0	-8.5	0.040
2.22	4.977	9.668	-0.152	0.0065	5.130	9.805	-0.072	5	63.9	17.128	-0.077	9.010	d	is	1	14	EBA-1 569.2 drillhole EBA-1, individual sample from depth 569.2, studied by XRD		0.86	132	0.8	0.4	-0.002
2.32	4.910	9.601	-0.184	0.015	5.063	9.738	-0.103	2	81.5	11.352	-0.122	5.900	d	is	1	14	EBA-1 1093 m, Lower Timeball Hill Fm, S Africa; immediately above Huronian glacials, studied by XRD	0.184	0.41	35	-15.5	-8.2	0.039
2.32	5.633	11.059	-0.234	0.023	5.783	11.182	-0.149	2	79.9	13.984	-0.202	7.216	d	is	1	14	EBA-1 1151.7m, Lower Timeball Hill Fm, S Africa; immediately above Huronian glacials. studied by XRD	0.188	0.35	7	-19.4	-10.2	0.049
2.32	4.203	8.264	-0.181	0.002	4.359	8.413	-0.104	2	76.4	9.675	-0.134	4.998	d	is	1	14, 52	H-68-3, 3410m, Eplett drillhole, Gowganda Fm, ON, Canada, above all Huronian glacials, studied for XRD	0.142	0.395	23	-19.7	-10.4	0.049
2.4	4.535	8.850	-0.160	0.003	4.690	8.994	-0.082	2	72.8	11.370	-0.095	5.936	d	is	1	14, 12	Drillcore A-77-10, 441 m, Pecors Fm (379-678'), L Huronian Supergroup, ON, Canada, studied by XRD	0.137	0.203	52	-9.6	-5.1	0.024
2.4	4.469	8.720	-0.157	0.015	4.625	8.866	-0.079	4	74.5	10.809	-0.088	5.646	d	is	1	14	Drillcore 150-4, 4912 m, Pecors Fm, L Huronian Supergroup, Canada, above the oldest Huronian glacial		0.3	63	-9.4	-5.0	0.024
2.4	4.468	8.712	-0.154	0.018	4.624	8.858	-0.075	14	74.5	10.792	-0.081	5.644	d	is	1	14	Drillcore 150-4, Pecors Fm, Lower Huronian Supergroup, Canada, above the oldest Huronian glacial		0.3	71	-8.0	-4.2	0.020
2.44	4.705	9.179	-0.163	0.020	4.860	9.321	-0.085	8	81.6	10.678	-0.094	5.570	d	is	1	14, 12	Drillcore DD04, 98.8m, Turee Creek Fm, Hamersley, W Australia, above oldest glacial, studied by XRD	0.151	0.23	53	-10.5	-5.5	0.026
2.44	4.393	8.603	-0.171	0.001	4.549	8.750	-0.093	2	73.4	10.739	-0.120	5.577	d	cs	10	14, 12	Pecors and McKim Fm, L Huronian, Canada, drill cores A-77-10, A-77-9 ,143-4,150-4, below all glacials	0.120	0.3	32	-15.3	-8.1	0.038
2.5	6.250	12.024	-0.129	0.003	6.396	12.136	-0.042	2	*75*	17.271	-0.014	9.148	d	cs	12	14, 53	Mt McRae Shale, Hamersley Basin, W Australia, drillcores WLT-02, WLT-10, secondary carbonate present	0.209	0.3	51	-4.9	-2.6	0.012
2.75	4.819	9.300	-0.115	0.018	4.973	9.441	-0.036	2	84.4	10.553	-0.020	5.578	d	cs	16	14, 53	Manjeri Fm, Bubi and Gweru Greenstone Belts, Zimbabwe, drillcores 690B92-02,696C92-01,711B93-02	0.191	0.3	134	-1.0	-0.5	0.003
2.95	3.960	7.757	-0.156	0.034	4.117	7.911	-0.079	3	82.6	8.397	-0.084	4.370	d	cs	3	55	Monzaan Group, Pongola Supergroup, South Africa, drillcore	0.143	0.3	32	-17.8	-9.4	0.044
2.95	3.860	7.491	-0.114	0.011	4.018	7.646	-0.038	4	82.7	7.988	-0.021	4.217	o	is	1	14,56,57	DRB-3 , W Rand Group, Government Subgroup, Witwatersrand Supergroup, S Africa, sample from mine	0.190	0.3	113	-5.7	-3.0	0.014
3.2	5.376	10.393	-0.138	0.014	5.527	10.523	-0.055	3	69.7	15.943	-0.033	8.425	d	cs	13	14, 58	Clutha Fm, Moodies Group, Barberton Greenstone Belt, S Africa, short drillcles, few % secondary silicification	0.172	0.3	136	4.5	2.4	-0.011
3.20	5.643	10.955	-0.156	0.015	5.793	11.079	-0.085	3	65.3	20.306	-0.119	10.653	d	is	1	14, 58	Drillcore 28-148, 43.3m, Clutha Fm, Moodies Gp, Barberton Greenstone Belt, S Africa, some sec. silicification	0.158	0.3	28	-6.8	-3.6	0.017
3.40	4.640	9.007	-0.138	0.011	4.795	9.150	-0.059	3	76.4	11.072	-0.049	5.825	o	is	1	14, 59	KS-1 , Kromberg Fm Onverwacht Group, Barberton Greenstone Belt, South Africa, minor sec. silicification	0.253	0.3	76	-7.1	-3.8	0.018
3.5	6.737	12.952	-0.134	0.006	6.880	13.052	-0.044	2	73.3	19.987	-0.015	10.588	o	cs	7	14, 60	Daitari Greenstone Belt, India, uncertain age, probably 3.5 Ga, some secondary silicification	0.168	0.3	58	-1.1	-0.6	0.003
3.75	5.779	11.117	-0.119	0.006	5.928	11.239	-0.035	3	*75*	15.479	0.001	8.212	o	is	1	61	sample R5310 , chloritized siltstone, surface outcrop, Isua greenstone belt, Greenland		0.01	79	0.6	0.3	-0.002
3.80	3.172	6.150	-0.091	0.014	3.333	6.315	-0.017	3	65.3	4.765	0.102	2.630	o	gr		62	Chinese granite -mylonitized, oldest rock in China 3.8Ga						
mantle	2.545	5.003	-0.121	0.034	2.677	5.140	-0.049										SCO average (n=12)						
mantle	2.681	5.153	-0.053	0.007	-	-	-										SCO analyzed by Pack et al., 2016 Rapid Comm Mass Spectrometry, 30, 1495.						
mantle	2.690	5.280	-0.101	0.015													SCO analyzed Pack and Herwartz, 2014 , EPSL 390, 138.						

Details of the shale samples are taken from refs [12,14,36–64].

*Normalized to the mantle value, −0.05‰ $\Delta^{17}O$.

†Data from ref. [8].

‡CIA is calculated using X-ray fluorescence (XRF) analysis of rocks where available. A generically average CIA of 75 is assumed for samples without chemical analysis or for which inductively coupled plasma mass spectrometry (ICPMS) analysis shows the calcium carbonate content to be greater than 5 wt% (most likely because of to secondary carbonate). A CIA of 100 is assumed for samples of 100% clay.

¶Corrected using the $\delta'^{18}O$ and $\Delta'^{17}O$ values of igneous crust, namely 7‰ and −0.07‰ respectively.

36. Ovey, C. D. Preliminary results from the study of an ocean core obtained by the Swedish Deep-Sea Expedition, 1947-48. *J. Glaciol.* **1**, 370–373 (1950).
37. Kah, L. C. et al. $\delta^{13}C$ stratigraphy of the Proterozoic Bylot Supergroup, Baffin Island, Canada: implications for regional lithostratigraphic correlations. *Can. J. Earth Sci.* **36**, 313–332 (1999).
38. Campbell, A. S. & Clark, B. L. Miocene radiolarian faunas from southern California. *Geol. Soc. Am. Spec. Pap.* **51**, 1–78 (1944).
39. Retallack, G. J. Lateritization and bauxitization events. *Econ. Geol.* **105**, 665–667 (2010).
40. Ryu, I.-C. Petrography, diagenesis and provenance of Eocene Tyee Basin sandstones, southern Oregon Coast Range: new view from sequence stratigraphy. *Isl. Arc* **12**, 398–410 (2003).
41. McKee, E. H. & Gangloff, R. A. Stratigraphic distribution of archaeocyathids in the Silver Peak Range and the White and Inyo Mountains, western Nevada and eastern California. *J. Paleontol.* **43**, 716–726 (1969).
42. Pecoits, E. et al. U–Pb detrital zircon ages from some Neoproterozoic successions of Uruguay: provenance, stratigraphy and tectonic evolution. *J. S. Am. Earth Sci.* **71**, 108–130 (2016).
43. Thomson, D., Rainbird, R. H., Planavsky, N., Lyons, T. W. & Bekker, A. Chemostratigraphy of the Shaler Supergroup, Victoria Island, NW Canada: a record of ocean composition prior to the Cryogenian glaciations. *Precambr. Res.* **263**, 232–245 (2015).
44. Key, R. M. et al. The western arm of the Lufilian Arc in NW Zambia and its potential for copper mineralization. *J. Afr. Earth Sci.* **33**, 503–528 (2001).
45. Podkovyrov, V. N. et al. Provenance and source rocks of Riphean sandstones in the Uchur-Maya region (east Siberia): implications of geochemical data and Sm-Nd isotopic systematics. *Stratigr. Geol. Correl.* **15**, 41–56 (2007).
46. Cullers, R. L. & Podkovyrov, V. N. The source and origin of terrigenous sedimentary rocks in the Mesoproterozoic Ui group, southeastern Russia. *Precambr. Res.* **117**, 157–183 (2002).
47. Vinogradov, V. I., Veis, A. F., Bujakaite, M. I. & Golovin, D. I. Isotopic evidences of epigenetic transformations and the problem of the age of Precambrian rocks in the Yudoma-Maya trough, eastern Siberia. *Lithol. Miner. Resour.* **35**, 141–152 (2000).
48. Addison, W. D., Brumpton, G. R., Davis, D. W., Fralick, P. W. & Kissin, S. A. Debrisites from the Sudbury impact event in Ontario, north of Lake Superior, and a new age constraint: are they base-surge deposits or tsunami deposits? *Geol. Soc. Am. Spec. Pap.* **465**, 245–268 (2010).
49. Sheppard, S. et al. A new Paleoproterozoic tectonic history of the eastern Capricorn Orogen, Western Australia, revealed by U-Pb zircon dating of micro-tuffs. *Precambr. Res.* **286**, 1–19 (2016).
50. Kipp, M. A., Stüeken, E. E., Bekker, A. & Buick, R. Selenium isotopes record extensive marine suboxia during the Great Oxidation Event. *Proc. Natl Acad. Sci. USA* **114**, 875–880 (2017).
51. Rouxel, O. J., Bekker, A. & Edwards, K. J. Iron isotope constraints on the Archean and Paleoproterozoic ocean redox state. *Science* **307**, 1088–1091 (2005).
52. Rasmussen, B., Bekker, A. & Fletcher, I. R. Correlation of Paleoproterozoic glaciations based on U–Pb zircon ages for tuff beds in the Transvaal and Huronian Supergroups. *Earth Planet. Sci. Lett.* **382**, 173–180 (2013).
53. Rasmussen, B., Blake, T. S. & Fletcher, I. R. U–Pb zircon age constraints on the Hamersley spherule beds: evidence for a single 2.63 Ga Jeerinah-Carawine impact ejecta layer. *Geology* **33**, 725–728 (2005).
54. Hofmann, A., Bolhar, R., Dirks, P. & Jelsma, H. The geochemistry of Archaean shales derived from a mafic volcanic sequence, Belingwe greenstone belt, Zimbabwe: provenance, source area unroofing and submarine versus subaerial weathering. *Geochim. Cosmochim. Acta* **67**, 421–440 (2003).
55. Planavsky, N. J. et al. Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. *Nat. Geosci.* **7**, 283–286 (2014).
56. Kositcin, N. & Krapež, B. Relationship between detrital zircon age-spectra and the tectonic evolution of the Late Archaean Witwatersrand Basin, South Africa. *Precambr. Res.* **129**, 141–168 (2004).
57. Armstrong, R. A., Compston, W., Retief, E. A., Williams, I. S. & Welke, H. J. Zircon ion microprobe studies bearing on the age and evolution of the Witwatersrand triad. *Precambr. Res.* **53**, 243–266 (1991).
58. Javaux, E. J., Marshall, C. P. & Bekker, A. Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits. *Nature* **463**, 934–938 (2010).
59. Byerly, G. R., Kröner, A., Lowe, D. R. & Todt, W. Prolonged magmatism and time constraints for sediment deposition in the early Archean Barberton greenstone belt: evidence from the Upper Onverwacht and Fig Tree groups. *Precambr. Res.* **78**, 125–138 (1996).
60. Mukhopadhyay, J. et al. Dating the oldest greenstone in India: a 3.51-Ga precise U-Pb SHRIMP zircon age for dacitic lava of the Southern Iron Ore Group, Singhbhum Craton. *J. Geol.* **116**, 449–461 (2008).
61. Myers, J. S. Protoliths of the 3.8–3.7 Ga Isua greenstone belt, West Greenland. *Precambr. Res.* **105**, 129–141 (2001).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

62. Liu, D. Y., Nutman, A. P., Compston, W., Wu, J. S. & Shen, Q. H. Remnants of ≥3800 Ma crust in the Chinese part of the Sino-Korean craton. *Geology* **20**, 339–342 (1992).
63. Retallack, G. J. & Krull, E. S. Neogene paleosols in the Sirius Group, Dominion, Antarctica. *US Antarctic J.* **30**, 10–14 (1997).
64. Bell, M. A. & Haglund, T. R. Fine-scale temporal variation of the Miocene stickleback *Gasterosteus doryssus*. *Paleobiology* **8**, 282–292 (1982).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Quantitative XRD data for a selection of studied shales

Sample		Age, Ga	SiO_2 norm	CIA	$\delta^{18}O$, ‰* published	$\delta^{18}O$, ‰, new	$\Delta^{17}O$, ‰, av	Quartz, %		s.d	Tot Mica, %	Albite, %		s.d	Total	Other
SC-93_1906		1.47	**64.00**	78.7	14.30	14.40	-0.198	10.6	±	2.9	88.4		±		99.0	
EBA-1_569.2		2.22	**86.30**	63.9	9.50	9.67	-0.152	70.1	±	4.3	11	14.4	±	1.2	95.5	
EBA-1_1093		2.32	**61.00**	81.5	10.10	9.60	-0.184	38.5	±	2.5	55.9		±		94.4	Amphibole (5.7%)
EBA-1_1151.7		2.32	**66.90**	79.9	11.74	11.06	-0.234	32.5	±	1.5	60.6		±		93.1	Amphibole (5.7%)
H-68-3_3410		2.32	**60.10**	76.4	8.34	8.26	-0.181	38.3	±	1.2	58.7	3	±	0.4	100.0	
A-77-10_441		2.40	**59.00**	72.8	9.47	8.85	-0.160	20.3	±	2.3	74	5.0	±	1.1	99.3	
DD04_98.8		2.46	**61.50**	81.6	8.52	9.18	-0.163	22.2	±	1.0	74.3	3.3	±	0.3	99.8	
89-MC-1_795		1.85	**61.50**	70.5	14.50			23.8	±	1.3	62.9	10.3	±	1.0	97.0	
EBA-1_574.7		2.32	**66.00**	71.3	7.79			40.6	±	2.8	49.3	10.1	±	6	100.0	
EBA-1_803.7		2.32	**63.10**	77.1	7.37			38.8	±	1	61.2		±		100.0	
DD04_112.5		2.47	**58.80**	82.0	9.22			23.8	±	0.6	76	0.3	±	0	100.1	

*Run as CO_2 for $\delta^{18}O$ only; previously published[14].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 3 | New XRF analyses of samples studied here

	Daitari Greenstone Belt, India			W Rand Group, Witwatersrand Supergroup S Africa	Barberton Greenstone Belt, Onverwacht Group, Kromberg Formation, South Africa	Bylot Supergroup, northern Baffin and Bylot Islands, Canada				Mwashya Subgroup, Upper Roan Group, Zambia					
Sample	DM 23	DM14	DM2	DRB3	KS-1	JD79 183B	JD79 297-1	JD79 I29 I-1	JD79	MJ 37	MJ 39	MJ 40	MJ 43	MJ 49	MJ
core/outcrop	o	o	o	mine	o	o	o	o	c	c	c	c	c	c	c
Age	3.5	3.5	3.5	2.95	3.40	1.27	1.27	1.27	1.13	0.75	0.75	0.75	0.75	0.75	0.75
n	1	1	1	1	1	1	1	1	3	1	1	1	1	1	4
SiO2	92.65	96.72	74.54	72.40	75.19	61.76	67.41	74.74	67.97	61.58	64.90	59.81	61.17	59.78	61.45
TiO2	0.16	0.08	0.74	0.46	0.75	0.63	1.47	0.67	0.92	1.69	1.21	0.83	1.10	1.51	1.27
Al2O3	5.17	1.77	16.80	11.43	16.96	12.77	18.38	15.33	15.50	16.76	17.81	19.15	11.50	14.74	15.99
Fe2O3	0.23	0.42	2.43	9.96	1.84	4.47	3.42	4.58	4.15	10.14	7.62	14.20	17.54	8.67	11.63
MnO	0.00	0.05	0.00	0.10	0.00	0.02	0.05	0.04	3.52	0.09	0.06	0.10	0.08	0.05	2.59
MgO	0.14	0.19	0.32	3.53	0.28	7.89	1.63	1.03	0.04	2.30	2.02	2.75	2.84	3.04	0.08
CaO	0.05	0.04	0.00	0.29	0.11	*5.47**	0.40	0.09	0.24	1.15	0.45	0.39	1.33	*5.03**	0.83
Na2O	0.00	0.10	0.00	0.00	0.00	1.72	3.08	0.05	1.62	2.70	2.18	0.26	1.85	3.16	2.03
K2O	1.53	0.54	5.00	1.61	4.65	4.84	3.89	3.25	3.99	3.24	3.43	2.27	2.33	3.57	2.97
P2O5	0.02	0.02	0.04	0.03	0.04	0.19	0.07	0.03	0.09	0.13	0.10	0.09	0.11	0.21	0.13
Total	99.96	99.94	99.87	99.81	99.82	94.29	99.79	99.80	97.96	99.78	99.78	99.85	99.85	94.75	98.80
CIA	74.74	69.60	75.60	82.70	76.60	*41.48*	65.14	81.62	62.75	63.16	69.40	85.18	59.83	*44.7*	64.46
ppm															
Rb	38	24	140	66	116	126	156	134	139	181	126	120	150	159	147
Sr			18	3	7	140	89	32	87	164	87	41	98	265	131
Ba	35		345	273	298	343	501	279	374	375	541	229	188	403	347
Zr	51	17	168	101	177	146	230	588	321	300	247	142	181	247	223
Y			17	12	13	14		25	19	30	33	10	22	29	25
Nb	9	6	13	10	11	14	42	18	24	52	44	22	30	37	37
Cs								4				27	18	4	
Sc	3	3	14	8	11	14	21	9	15	27	21	16	8	23	19
V	21	15	94	88	137	464	179	70	237	156	142	109	99	166	134
Cr	46	247	78	416	206	67	122	55	81	114	93	98	46	76	85
Ni	10	77	18	276	77	52	12	24	29	24	20	36		26	26
Cu		5		42	94	56		9	33		2	33	2	17	14
Zn		34	11	77	153	319	87	101	169	123	172	119	168	37	124
Ga	4		17	13	20	15	20	16	17	24	22	20	15	21	21
La	25		25	32	23	27	44	41	37	70	51	61	51	66	60
Ce	59	22	39	42	18	67	120	78	88	96	106	106	122	119	110
Pr	8	1	5	8	4	14	14	8	12	11	12	14	12	11	12
Nd	42		11	6		11	37	27	25	27	37	37	28	37	33
Hf					2										
Pb				1	18	5		31	18	13	1				7
Th				4		1		14	7	3	10			15	9
Ta									4						

Sample	Ui Fm Upper part of the Ui Group, Lakhanda Series, Uchur-Maya Region, Siberia, Russia	SC-93 Newland Formation, Lower Belt Supergroup Montana, USA	89MC Rove Formation, near Thunder Bay, Ontario, Canada	KDD Maraloo Formation, SE Capricorn Orogen, Australia	DH-51XX Zaonega Formation, lower Ludikovian Series, Karelia, Russia	STRAT-2 Sengoma Argillite Formation, Pretoria Series, Lobatse, Botswana	EBA-1 Upper Timeball Hill Fm, shale+siltstone, Potchefstroom Synclinorium, South Africa; above all glacials	EBA-1 and H-68 Eplett drillhole, Gowganda Formation, Huronian Supergroup, Cobalt area ON, Canada, above all Huronian glacials	A77 and 150 Pecors Formation (14'-292'), Lower Huronian Supergroup ON, Canada, above the oldest Huronian glacial	DD04 Turee Creek Group, Hamersley Province, Western Australia, above the oldest Huronian glacial	143/4 McKim Formation (2162-2550'), Lower Huronian Supergroup ON, Canada, below all glacials	WLT Mt McRae Shale, Hamersley Province, Western Australia	FVG Roy Hill Shale, Jeerinah Fm, Western Australia	Dominion Reef, South Africa	28 166 Clutha Formation, Moodies Group, Barberton Greenstone Belt, South Africa	Barberton Greenstone Belt, Onverwacht Group, Kromberg Formation, South Africa	DM, Daitari Greenstone Belt, India, slate
core/outcrop	o	c	c	c	c	c	c	c	c	c	c	c	c	c	c	o	o
Age	0.98	1.47	1.85	1.87	2.05	2.15	2.22	2.32	2.4	2.46	2.47	2.5	2.66	2.95	3.2	3.4	3.5
n	5	2	2	3	4	3	3	3	7	3	4	4	4	1	4	1	3
SiO2	59.96	63.7	56.2	64.34	64.13	66.12	70.53	64.54	58.22	60.05	58.26	46.18	64.29	72.40	60.83	75.19	74.54
TiO2	1.04	0.8	0.9	0.87	0.68	0.75	0.59	0.73	0.86	0.62	0.87	0.43	0.67	0.46	0.56	0.75	0.33
Al2O3	15.66	20.8	22.6	18.55	16.88	17.10	16.29	20.80	24.13	18.09	23.70	12.77	15.76	11.43	15.01	16.96	7.92
Fe2O3	9.96	8.2	7.0	7.21	4.20	6.11	6.07	6.71	5.67	9.90	6.20	4.39	6.10	9.96	7.84	1.84	1.03
MgO	2.45	1.3	2.8	3.74	2.40	2.84	1.09	1.64	2.31	7.36	2.09	10.29	3.40	3.53	5.08	0.28	0.22
MnO	0.05	0.0	0.0	0.63	0.05	0.02	0.02	0.20	0.04	0.04	0.04	0.41	0.08	0.10	0.17	0.00	0.02
CaO	0.41	0.1	0.4	0.41	2.77	0.97	0.70	0.59	0.25	0.16	0.22	3.56	1.40	0.29	*0.62*	0.11	0.03
K2O	1.88	4.7	5.4	3.03	1.08	3.60	3.56	3.77	7.52	3.46	7.58	7.67	7.11	0.00	5.44	0.00	0.03
Na2O	4.39	0.2	0.8	0.86	7.50	2.23	0.93	0.70	0.61	0.05	0.63	0.16	0.22	1.61	1.67	4.65	2.36
P2O5	0.46	0.1	0.1	0.10	0.09	0.11	0.09	0.14	0.11	0.12	0.12	0.08	0.08	0.03	0.10	0.04	0.03
Total	99.40	99.8	96.2	99.74	99.79	99.86	99.86	99.82	99.72	99.84	99.69	85.93	98.76	99.81	99.37	99.82	99.92
LOI		0.2	3.8	0.26	0.22	0.14	0.14	0.18	0.28	0.16	0.31	0.00	1.24		0.63		
CIA	74.80	78.8	74.3	76.49	56.60	65.45	70.48	77.07	71.95	82.22	70.82		60.72	82.70	*62.70*	76.60	73.31
ppm																	
Rb	193	129	225	148	37	184	183	181	329	140	320	224	151	66	142	116	67
Sr	29	71	87	36	202	88	84	89	44	9	36	60	32	3	247	7	18
Ba	563	337	1182	1805	662	857	422	626	1302	412	1151	377	406	273	712	298	190
Zr	246	230	159	137	123	164	176	157	146	138	156	108	247	101	121	177	78
Y		29	36	37	20	32	30	30	28	19	26	23	53	12	21	13	17
Nb	18	17	16	16	11	14	15	17	12	10	13	12	23	10	10	11	9
Sc	20	12	19	25	14	17	11	15	18	20	19	24	15	8	14	11	7
V		114	222	196	88	233	102	137	168	123	167	75	109	88	107	137	43
Cr	92	105	168	179	126	140	96	147	164	151	159	98	188	416	658	206	123
Ni		53	37	181	77	117	41	64	72	76	89	84	118	276	291	77	35
Cu		14	33	58	10	94	32	47	22	3	63	24	121	42	31	94	
Zn		49	163	187	55	170	94	102	64	70	56	355	96	77	84	153	22
Ga		25	31	24	20	24	20	26	31	24	30	17	25	13	19	20	11
La	54	42	53	95	55	34	53	51	43	52	43	39	54	32	35	23	25
Ce	89	88	108	200	119	74	116	110	91	93	88	85	119	42	68	18	40
Pr		9	11	21	13	9	13	13	9	11	9	11	11	8	7	4	5
Nd	37	41	40	62	39	23	52	45	37	43	39	26	51	6	23		26
Hf	7	7	7	5	4	5	6	6	5	5	5	4	6		3	2	
Pb		8	14	6		16	18	75	15	5	15	20	14	1	16	18	
Th	16	16	23	20	10	22	19	22	18	16	22	17	14	4	7		
Ta	6.15																

bdl, below detection limit; LOI, loss on ignition; p.p.m., parts per million.
*Secondary carbonate present.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0071-9

Large potential reduction in economic damages under UN mitigation targets

Marshall Burke[1,2,3]*, W. Matthew Davis[2] & Noah S. Diffenbaugh[1,4]

International climate change agreements typically specify global warming thresholds as policy targets[1], but the relative economic benefits of achieving these temperature targets remain poorly understood[2,3]. Uncertainties include the spatial pattern of temperature change, how global and regional economic output will respond to these changes in temperature, and the willingness of societies to trade present for future consumption. Here we combine historical evidence[4] with national-level climate[5] and socioeconomic[6] projections to quantify the economic damages associated with the United Nations (UN) targets of 1.5 °C and 2 °C global warming, and those associated with current UN national-level mitigation commitments (which together approach 3 °C warming[7]). We find that by the end of this century, there is a more than 75% chance that limiting warming to 1.5 °C would reduce economic damages relative to 2 °C, and a more than 60% chance that the accumulated global benefits will exceed US$20 trillion under a 3% discount rate (2010 US dollars). We also estimate that 71% of countries—representing 90% of the global population—have a more than 75% chance of experiencing reduced economic damages at 1.5 °C, with poorer countries benefiting most. Our results could understate the benefits of limiting warming to 1.5 °C if unprecedented extreme outcomes, such as large-scale sea level rise[8], occur for warming of 2 °C but not for warming of 1.5 °C. Inclusion of other unquantified sources of uncertainty, such as uncertainty in secular growth rates beyond that contained in existing socioeconomic scenarios, could also result in less precise impact estimates. We find considerably greater reductions in global economic output beyond 2 °C. Relative to a world that did not warm beyond 2000–2010 levels, we project 15%–25% reductions in per capita output by 2100 for the 2.5–3 °C of global warming implied by current national commitments[7], and reductions of more than 30% for 4 °C warming. Our results therefore suggest that achieving the 1.5 °C target is likely to reduce aggregate damages and lessen global inequality, and that failing to meet the 2 °C target is likely to increase economic damages substantially.

Anticipating the potential impacts of climate change is central to planning appropriate policy responses, including how to allocate resources among mitigation and adaptation options. By committing the international community to holding global warming to "well below 2 °C above pre-industrial levels" and pursuing a 1.5 °C target[1], the UN Paris Agreement increased the need for quantitative analysis of uncertainties in the costs and benefits of achieving highly resolved warming targets. In particular, because mitigation costs are thought to rise rapidly for more stringent targets[9], understanding the value of avoided impacts (what we term 'benefits') is central to evaluating the 1.5 °C target. Quantification of these potential benefits and their uncertainties is needed at the aggregate global level to guide coordinated global policy, as well as at a more local level to understand the distributional impacts of global policy choices[10]. Further, because the current national commitments imply warming[7] of 2.5–3 °C, quantifying the impact of exceeding the 1.5 °C and 2 °C targets is also critical to understanding the implications of policy choices.

Here we estimate the global and country-specific economic impacts of limiting warming to 1.5 °C relative to 2 °C, as well as the global impacts of projected warming under current mitigation commitments, separate from any mitigation costs incurred in achieving those targets. We measure potential global and country-level damages using gross domestic product (GDP), the total value of goods and services produced in a country in a given year. GDP is clearly an incomplete summary of the benefits of mitigation, and it cannot easily diagnose many sector-specific impacts (for example, in crop agriculture versus manufacturing). However, it does capture how sector-specific impacts interact and aggregate—a traditional challenge for sector-specific empirical work and model-based approaches to aggregation[11]. GDP also remains highly relevant to policy discussions, and the level and uncertainty in GDP impacts associated with the UN temperature targets has not been formally quantified.

We construct a probabilistic framework (Fig. 1) that incorporates uncertainty in (1) the historical relationship between temperature variability and economic growth, (2) the spatial pattern of future mean annual temperature change associated with a given level of aggregate emissions, (3) the future rate and pattern of economic development absent climate change, and (4) how future damages should be discounted.

To estimate the historical relationship between temperature and GDP, we use annual measurements of average temperature and growth in GDP per capita from 165 countries over the years 1960–2010. Following Burke et al.[4], we use a fixed-effects estimator that isolates the effect of temperature fluctuations from other time-invariant and time-varying factors that might be correlated with both temperature and economic output, and we estimate nonlinear response functions that allow the marginal effect of warming to differ as a function of countries' average temperatures. To quantify uncertainty in this historical relationship, we employ multiple bootstrapping approaches, estimating a separate response function for each re-sample (see Methods).

All estimated response functions relating GDP growth to temperature display a similar concave shape (Fig. 1a), suggesting that additional warming accelerates growth in cooler regions and slows growth in warmer regions. These findings are consistent with a large body of work demonstrating nonlinear responses of economic outcomes to changes in temperature[12–17]. However, there is uncertainty in the temperature at which additional warming begins to generate damages rather than benefits (the 'temperature optimum'), with a median estimate of 13.1 °C but a 5%–95% range of 9.7–16.8 °C. Because much of today's GDP is produced in areas just beyond the median estimated optimal temperature (density plot, bottom of Fig. 1a), uncertainty in this optimum leads to substantial overall uncertainty in both the magnitude and sign of the impact of additional warming.

We project impacts under different levels of future warming by combining these historical response functions with the Intergovernmental Panel on Climate Change (IPCC) projections of future climate[18]. The climate model experiments used by the IPCC involve dozens of general circulation models (GCMs) run under four forcing pathways (called

[1]Department of Earth System Science, Stanford University, Stanford, CA, USA. [2]Center on Food Security and the Environment, Stanford University, Stanford, CA, USA. [3]National Bureau of Economic Research, Cambridge, MA, USA. [4]Woods Institute for the Environment, Stanford University, Stanford, CA, USA. *e-mail: mburke@stanford.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

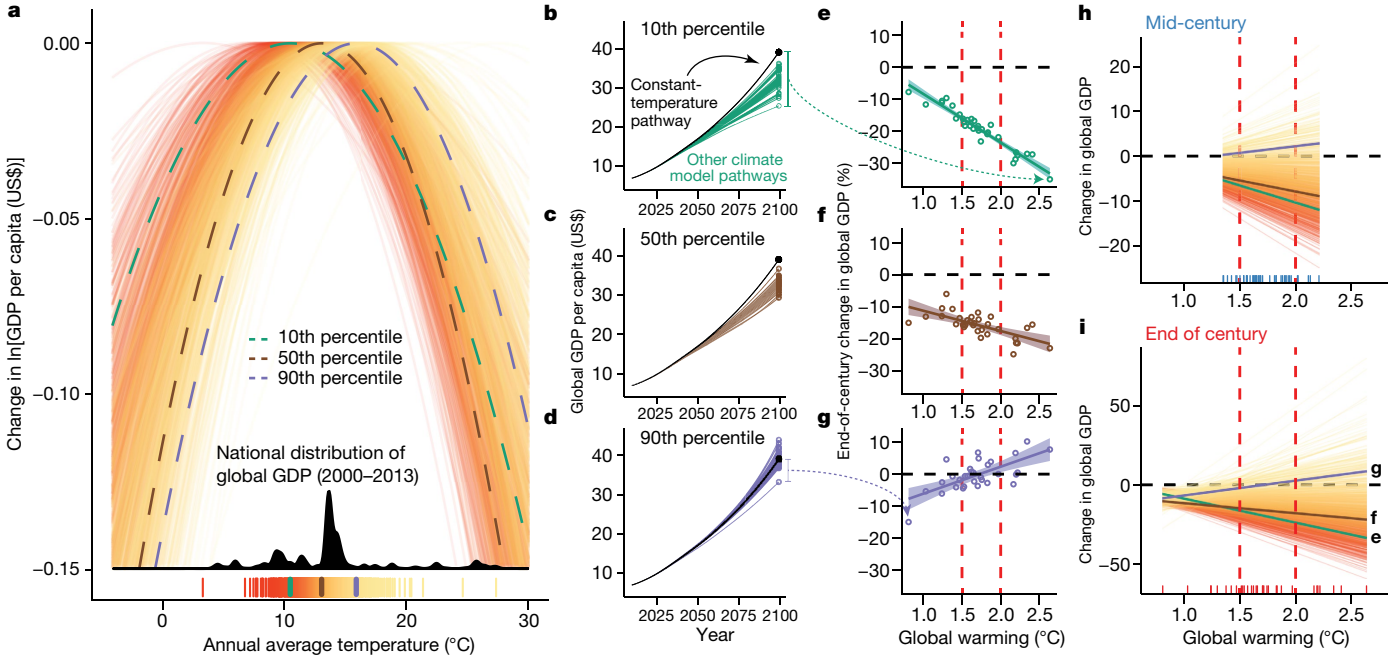


Fig. 1 | Deriving impact projections. a, Historical response of per capita GDP growth rates to temperature. Each curve is the response function estimated from one of 1,000 bootstraps of a historical regression with colour corresponding to the temperature at which it optimizes (redder colours for cooler optima). The green, brown and purple dashed curves highlight bootstraps at the 10th, 50th and 90th percentiles of optimizing temperatures, respectively. The rug plot at the bottom shows the distribution of optimizing temperatures across bootstraps using the same colour scheme. The density plot in black shows the GDP-weighted distribution of baseline average national temperatures. **b**–**d**, Projected future economic pathways under different historical response functions. Black lines represent the pathway of global GDP per capita, assuming no future warming. Coloured lines are pathways corresponding to the response functions at the 10th, 50th and 90th percentiles highlighted in **a**, under warming projections from 32 GCMs consistent with RCP2.6. Points represent values projected for 2099. **e**–**g**, Projected climate impact on global GDP per capita by 2099 for the same response functions, equivalent to the percentage difference between the black points and coloured points in **b**–**d**. The warming on the *x* axes is the global warming projected for 2099 by GCMs running RCP2.6, relative to a pre-industrial benchmark. Red vertical dashed lines mark 1.5 °C and 2.0 °C warming. Linear ordinary least-squares models are fitted for each of the response functions, with the slope estimating the per-degree impact of global warming on global GDP per capita. Shaded areas represent the 95% confidence interval of the ordinary least-squares fit. **i**, The linear fits from **e**–**g**, but for all bootstrapped response functions instead of just the three highlighted in **b**–**g**. The colours correspond to the optimizing temperatures of the response functions, as in **a**. The rug plot at the bottom marks global warming for the end of the century (2099) projected by the 32 GCMs consistent with RCP2.6, equivalent to the *x*-axis values of points in **e**–**g**. **h**, Equivalent to **i** but for mid-century (2049) projections based on 42 GCMs consistent with RCP4.5.

representative concentration pathways, or RCPs). Each GCM realization contains a temperature trajectory for each country and, in aggregate, for the globe. Because temperature affects both the level and the growth rate of economic output[4,11], and because growth effects compound over time, the projected differential impacts of 1.5 °C versus 2 °C are a function of the time horizon. We calculate differential impacts under the two targets using temperature changes for the mid-century (2046–2065) and end-of-century (2081–2100) periods used by the IPCC, focusing on output from those RCPs whose ensemble range spans 1.5 °C and 2 °C for a given time period (Methods). We use projections from the relevant shared socioeconomic pathways (SSPs) to define the secular evolution of population and economic development[6,19], (Fig. 1b–d, Extended Data Fig. 2).

Economic impacts are calculated relative to a constant-temperature counterfactual and are then aggregated globally (weighting by population), resulting in a unique estimate of global impact for each bootstrap–GCM–SSP–year combination. We present two measures of these relative impacts: the percentage difference in annual GDP at the end of the chosen projection period and the discounted present value of absolute GDP differences accumulated over that span. For the second measure we employ a range of discounting schemes, including fixed rates of 2.5%–5% per annum (where a 5% discount rate assumes that society values a given amount of consumption in one year roughly 5% less than it values it today) and time-varying rates that depend on the levels of and uncertainty in realized growth (Methods).

We estimate the benefits of 1.5 °C versus 2 °C by fitting a linear least-squares regression relating either measure of relative economic impact to the global warming projected by each GCM that archives the RCP (Fig. 1e–g). We repeat this procedure for every bootstrapped response function to arrive at a distribution of estimated impacts for the chosen combination of GCM, SSP and projection period. See Methods for a full derivation.

Most response functions generate more negative global impacts at 2 °C than at 1.5 °C (Fig. 1h–i, Extended Data Fig. 2). Cooler estimated historical optima (red colours) generate steeper negative responses to additional warming, implying greater benefits from more stringent mitigation. We estimate that limiting warming to 1.5 °C instead of 2 °C by mid-century would lead to an increase in global GDP of 1.5%–2.0% (median estimate; Fig. 2a) and US$7.7–11.1 trillion in discounted avoided damages under a 3% fixed annual discount rate. Meeting these targets at the end of the century is estimated to lead to median gains in global GDP per capita of 3.4% and discounted avoided damages of US$36.4 trillion.

We use the distributions of bootstrapped estimated impacts to quantify the probability that more stringent mitigation yields benefits of different magnitudes (Extended Data Table 1). We estimate that achieving the 1.5 °C target at mid-century (2046–2065) would lead to a 68%–76% chance of overall cumulative net benefit relative to 2 °C under a fixed 3% discount rate. Under the same discount rate, we estimate a 43%–53% chance of discounted cumulative benefits exceeding US$10 trillion and a 4%–8% chance of exceeding $30 trillion, which is about 40% of current global GDP. For the end of the century (2081–2100), we estimate a >75% chance of net gain in per capita global GDP, an approximately 38% chance that benefits exceed US$50 trillion, and

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

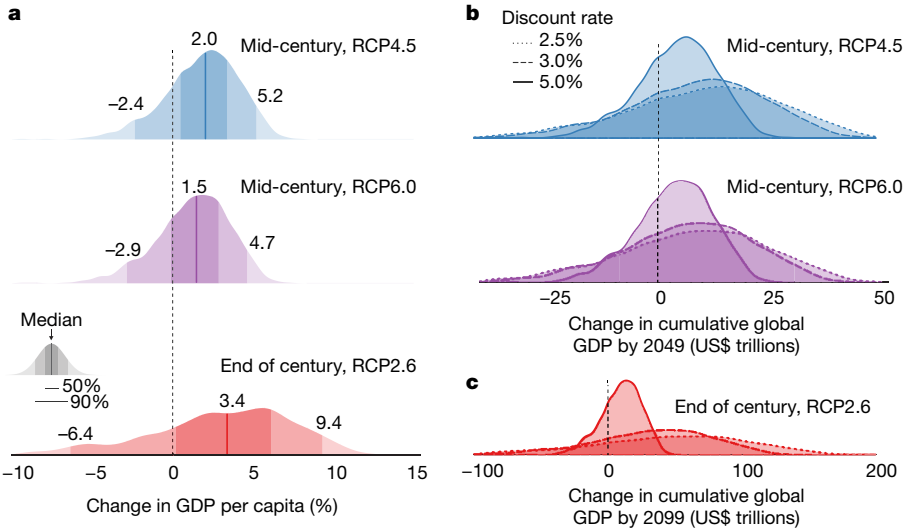


Fig. 2 | Global impact of limiting global warming to 1.5 °C relative to 2 °C. a, Probability distribution of the percentage change in global GDP per capita for 1.5 °C versus 2 °C by mid-century and by the end of the century, as derived from the slopes of the linear fits across response functions illustrated in Fig. 1h–i. Positive values indicate reduced damages at 1.5 °C of global warming as compared to 2 °C. Values above distribution report percentage changes at the 10th, 50th and 90th percentiles of distribution. **b**, Probability distribution of the change in cumulative global GDP by mid-century, assuming discount rates of 2.5% (dotted line), 3% (dashed line) and 5% (filled line). **c**, The equivalent for the end of the century.

an approximately 5% chance that benefits exceed US$100 trillion (3% discount rate; Extended Data Table 2).

While end-of-century estimates of the magnitude of absolute impacts are sensitive to choices about discounting (Extended Data Fig. 3, Extended Data Table 1), estimates of the probability of positive benefits are much less so (Extended Data Tables 2 and 3). Results are also relatively insensitive to alternative bootstrap resampling approaches, to different SSPs, and to alternative assumptions about the time path of future warming for a given RCP (Extended Data Figs. 4, 5). Inclusion of additional lags of temperature in the historical regression—a common

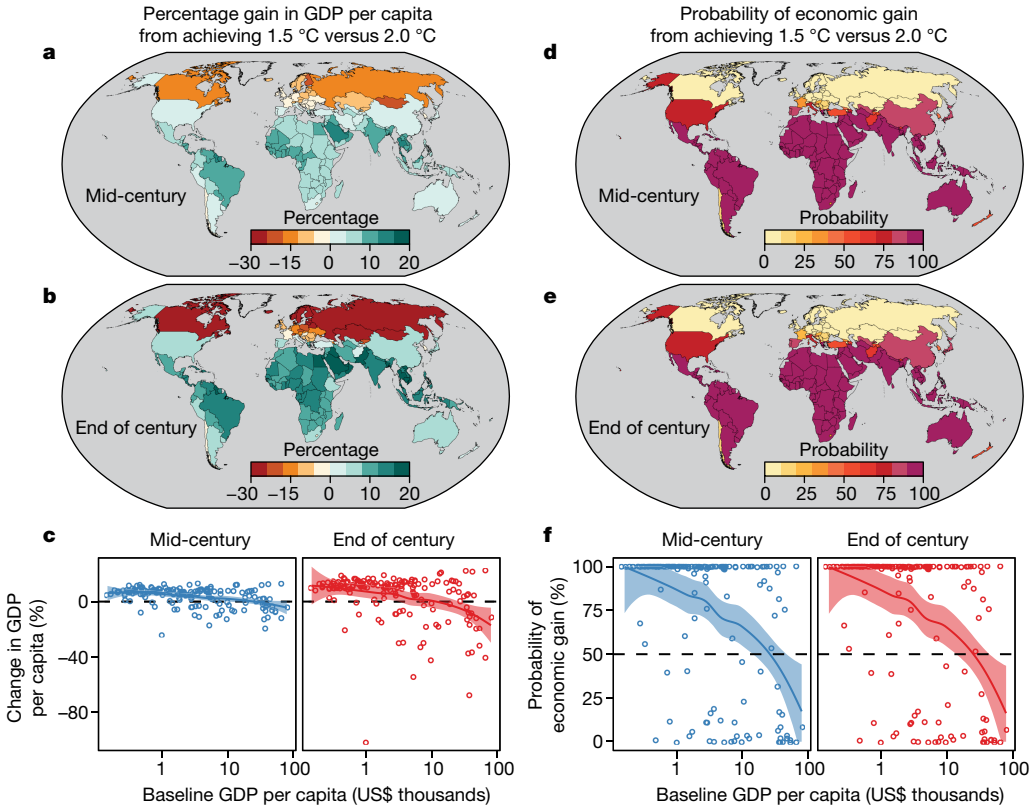


Fig. 3 | Country-level impact of limiting global warming to 1.5 °C relative to 2 °C. a, **b**, Median estimates of impacts on change in GDP per capita under 1.5 °C versus 2 °C, for mid-century and the end of the century. Positive values indicate reduced damages at 1.5 °C of global warming as compared to 2 °C. **c**, Median estimated impacts as a function of each country's baseline GDP per capita, with each country weighted equally. Lines represent local polynomial regression fits to the data with the corresponding 95% confidence intervals shaded in grey. **d**–**f**, As in **a**–**c**, but for the probability of per capita GDP gain, calculated as the percentage of bootstrap response functions projecting a net gain in a country's GDP per capita under 1.5 °C of global warming as compared to 2 °C.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

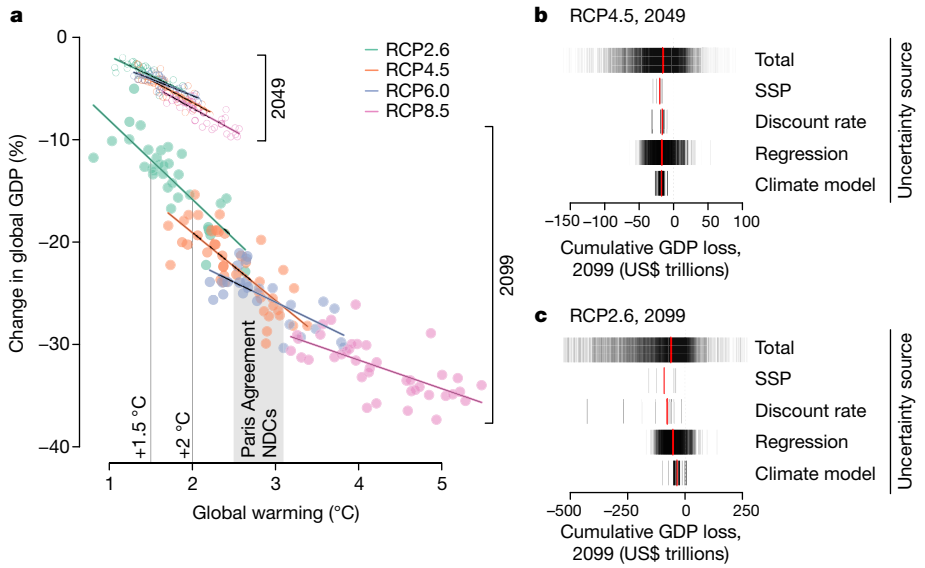


Fig. 4 | The impact of global warming on global GDP per capita, relative to a world without warming, for different forcing levels. a, Projected percentage change in global GDP for different climate models under different RCP forcing scenarios, relative to a no-warming baseline (median bootstrap, SSP1). Colours denote different RCPs. Unfilled points show mid-century projections, filled points show end-of-century projections. Vertical lines show the UN temperature targets as well as the range of estimates of end-of-century warming under current Paris commitments[7]. Warming is relative to pre-industrial levels. **b**, **c**, Sources of uncertainty in estimates of global warming on cumulative global GDP loss for a given forcing level. Total uncertainty in the impact of warming on global GDP under a given forcing scenario is a combination of uncertainty in how economies respond to warming ('historical regression uncertainty'), uncertainty across climate models in the amount and pattern of warming for a given level of forcing ('climate model uncertainty'), uncertainty in baseline future growth rates across baseline socioeconomic scenarios ('SSP uncertainty'), and plausible alternatives for how to specify the discount rate ('discount rate uncertainty'). Values show cumulative global GDP losses in trillions of US$ for mid-century under RCP4.5 (**b**) and the end of the century under RCP2.6 (**c**), either with all factors allowed to vary ('total uncertainty') or with the listed factor allowed to vary and all others fixed at their median (see Methods). Each vertical line is a point estimate; for example, with 32 climate models running RCP2.6 there are 32 estimates shown for 'climate model uncertainty' in **c**. Red lines are the median estimate across each uncertainty distribution.

approach to capturing persistent growth effects[11]—amplifies the effect of temperature on growth rates and results in larger estimates of benefits under 1.5 °C (Extended Data Fig. 4). Other potential sources of uncertainty, such as uncertainty in the secular rate of growth beyond the scenarios prescribed by the SSPs, were not quantified and could increase overall impact uncertainty.

At the country level, both the magnitude and the uncertainty of potential benefits are highly non-uniform. We find that 71% of countries—encompassing about 90% of projected global population—exhibit a >75% chance of experiencing positive economic benefits at 1.5 °C relative to 2 °C (Fig. 3), and 59% of countries exhibit a >99% chance. These countries include the three largest economies (the USA has a 76% chance of positive benefits; China 85%; Japan 81%) (Fig. 3, maps). They also include a large fraction of the world's poorest countries, with the likelihood of economic gains rising rapidly at lower levels of GDP per capita (Fig. 3c, f). Many of the countries that exhibit a high probability of economic benefits from 1.5 °C are concentrated in the tropics and sub-tropics, where both current and future temperatures are warmer than the economic optimum[4]. As a result, even small reductions in future warming in these countries can generate substantial increases in per capita GDP, with many countries in the tropics exhibiting per capita GDP 10%–20% higher at 1.5 °C than 2 °C by the end of the century (Fig. 3a, b, d, e). The opposite is true for a smaller number of high-latitude countries, where 1.5 °C is estimated to slow growth and generate a high probability of negative impacts relative to 2 °C. Achieving the 1.5 °C target will thus have unequal consequences, with today's poorest countries benefiting the most.

Despite the Paris Agreement's focus on the 1.5 °C and 2 °C targets, its actual Nationally Determined Contributions (NDCs) are instead consistent with 2.5–3 °C of global warming[7]. We estimate that this level of warming could lead to a reduction in global GDP as high as 10% by mid-century and 15%–25% by the end of the century (median estimates across SSPs; Fig. 4 and Extended Data Fig. 6), relative to a world that did not warm beyond 2000–2010 levels. In addition, failing to meet the NDC commitments is likely to lead to reductions in global GDP that exceed 25% by the end of the century. Uncertainty in these estimates is driven much more by uncertainty in economic parameters—namely, the economic response to warming and the discount rate—than by uncertainty in the pattern and magnitude of temperature change reflected in the climate model ensemble (Fig. 4b and c), highlighting the importance of better constraining these economic parameters[20].

Because our future impact estimates are based on observed historical economic responses to temperature variability, our projections will misstate impacts if the relationship between future annual temperatures and climatic extremes differs from what has occurred historically, or if future societies respond differently from societies in the recent past—although there is growing evidence that economic development might not fundamentally alter these economy–environment linkages[4,15–17]. We also cannot account for historically unprecedented changes, such as large-scale loss of land ice and associated sea level rise, which are more likely to occur[8,21] at 2 °C than 1.5 °C and are expected to exacerbate impacts[22,23].

To support policy decisions, our estimates of avoided damages need to be compared against the costs of meeting the UN targets. To our knowledge, no comparable estimates of global abatement costs through to the end of the century currently exist. However, a recent estimate[24] suggests that achieving emissions levels in 2030 that are consistent with the 1.5 °C target will lead to approximately US$300 billion in additional (non-discounted) abatement costs relative to emissions consistent with 2 °C. This estimate of abatement costs is >30 times smaller than our median estimate of (discounted) mid-century avoided damages.

Not accounting for abatement costs, our results suggest that 1.5 °C global warming is "likely"[25] to result in substantial economic benefits relative to 2 °C, with foregone damages probably in the tens of trillions of dollars and 59% of countries "virtually certain"[25] to benefit. Given that most of these countries feature large populations or high poverty rates

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

or both, our results suggest that achieving more stringent mitigation targets will probably generate a net global benefit, with particularly large benefits for the poorest populations.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0071-9.

Received: 2 August 2017; Accepted: 28 February 2018;
Published online 23 May 2018.

1. UNFCCC *Adoption of the Paris Agreement. I: Proposal by the President. Draft Decision CP.21* https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf (United Nations Office, Geneva, 2015).
2. Nature Editorial Board. The maximum climate ambition needs a firm research backing. *Nature* **537**, 585–586 (2016).
3. Hallegatte, S. et al. Mapping the climate change challenge. *Nat. Clim. Chang.* **6**, 663–668 (2016).
4. Burke, M., Hsiang, S. M. & Miguel, E. Global non-linear effect of temperature on economic production. *Nature* **527**, 235–239 (2015).
5. Collins, M. et al. in Climate Change 2013: the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (eds Stocker, T. et al.) Ch. 12 (Cambridge Univ. Press, Cambridge, 2014).
6. Riahi, K., van Vuuren, D. P. & Kriegler, E. The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: an overview. *Glob. Environ. Change* **42**, 153–168 (2017).
7. Rogelj, J. et al. Paris agreement climate proposals need a boost to keep warming well below 2 °C. *Nature* **534**, 631–639 (2016).
8. Schleussner, C.-F. et al. Science and policy characteristics of the Paris Agreement temperature goal. *Nat. Clim. Chang.* **6**, 827–835 (2016).
9. Rogelj, J., McCollum, D. L., Reisinger, A., Meinshausen, M. & Riahi, K. Probabilistic cost estimates for climate change mitigation. *Nature* **493**, 79–83 (2013).
10. Hallegatte, S., Fay, M., Bangalore, M., Kane, T. & Bonzanigo, L. *Shock Waves: Managing the Impacts of Climate Change on Poverty* (World Bank Publications, Washington, 2015).
11. Dell, M., Jones, B. F. & Olken, B. A. Temperature shocks and economic growth: evidence from the last half century. *Am. Econ. J. Macroecon.* **4**, 66–95 (2012).
12. Schlenker, W. & Roberts, M. J. Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proc. Natl Acad. Sci. USA* **106**, 15594–15598 (2009).
13. Hsiang, S. M. Temperatures and cyclones strongly associated with economic production in the Caribbean and central america. *Proc. Natl Acad. Sci. USA* **107**, 15367–15372 (2010).
14. Carleton, T. A. & Hsiang, S. M. Social and economic impacts of climate. *Science* **353**, aad9837 (2016).
15. Deryugina, T. & Hsiang, S. M. *Does the Environment Still Matter? Daily Temperature and Income in the United States*. NBER working paper 20750 (National Bureau of Economic Research, Cambridge, 2014); http://www.nber.org/papers/w20750.
16. Colacito, R., Hoffmann, B. & Phan, T. *Temperatures and Growth: A Panel Analysis of the United States*. IDB working paper IDB-WP-676 (Inter-American Development Bank, Washington, 2016); https://publications.iadb.org/handle/11319/7654.
17. Zhang, P., Deschenes, O., Meng, K. & Zhang, J. Temperature effects on productivity and factor reallocation: evidence from a half million Chinese manufacturing plants. *J. Environ. Econ. Manage.* **88**, 1–17 (2018).
18. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
19. O'Neill, B. C., Kriegler, E. & Ebi, K. L. The roads ahead: narratives for shared socioeconomic pathways describing world futures in the 21st century. *Glob. Environ. Change* **42**, 169–180 (2017).
20. Burke, M. et al. Opportunities for advances in climate change economics. *Science* **352**, 292–293 (2016).
21. Seneviratne, S. I., Donat, M. G., Pitman, A. J., Knutti, R. & Wilby, R. L. Allowable CO_2 emissions based on regional and impact-related climate targets. *Nature* **529**, 477–483 (2016).
22. Cai, Y., Lenton, T. M. & Lontzek, T. S. Risk of multiple interacting tipping points should encourage rapid CO_2 emission reduction. *Nat. Clim. Chang.* **6**, 520–525 (2016).
23. Hsiang, S. et al. Estimating economic damage from climate change in the United States. *Science* **356**, 1362–1369 (2017).
24. Hof, A. F. et al. Global and regional abatement costs of nationally determined contributions (NDCs) and of enhanced action to levels well below 2 °C and 1.5 °C. *Environ. Sci. Policy* **71**, 30–40 (2017).
25. Mastrandrea, M. D. et al. The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups. *Clim. Change* **108**, 675 (2011).

Acknowledgements We thank L. Goulder, S. Hsiang and D. Lobell for comments. We thank the Erol Foundation for funding.

Reviewer information *Nature* thanks R. Kopp and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Author contributions: M.B. and N.S.D. designed the research. M.B., W.M.D. and N.S.D. analysed data, interpreted results and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0071-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Deriving the historical response function. To understand the historical relationship between temperature and economic output, we assemble annual data on country-level GDP per capita from the World Bank's World Development Indicators, using data on 165 countries over the period 1960 to 2010. Growth is computed as the first difference of the natural logarithm of the annual purchasing power parity-adjusted per capita GDP series in each country. These data are then merged with temperature and precipitation data from the University of Delaware[26].The gridded monthly temperature and precipitation data are aggregated temporally to the annual level and spatially to the country level. We then follow ref. [4] and estimate a panel fixed effects model:

$$\Delta\log(y_{it}) = \beta_1 T_{it} + \beta_2 T_{it}^2 + \lambda_1 P_{it} + \lambda_2 P_{it}^2 + \mu_i + \upsilon_t + \theta_{1i} t + \theta_{2i} t^2 + \varepsilon_{it} \quad (1)$$

where y_{it} is per capita GDP in country i in year t, T and P are the average temperature and precipitation in year t, μ_i are country-fixed effects (dummies) that control for time-invariant differences between countries, υ_t are year-fixed effects that account for common global shocks in a given year, and $\theta_{1i}t + \theta_{2i}t^2$ are country-specific linear and quadratic time trends, which allow temperature and growth to evolve flexibly at the country level.

Equation (1) is estimated simultaneously on our global sample of country-years ($N = 6{,}584$). Point estimates for β_1 and β_2 are statistically significant in this regression ($\beta_1 = 0.0127$, standard error 0.0032, $P < 0.001$; $\beta_2 = -0.0005$, standard error 0.0001, $P < 0.001$).

Equation (1) assumes that there is a single response function (described by β_1 and β_2) that specifies the overall global relationship between income growth and changes in temperature, but that individual countries can respond differently to warming as a function of their average temperature (which can be seen by differentiating equation (1) with respect to temperature). Past work has shown that average temperature—rather than other correlated factors such as average income—is the main source of heterogeneity in how countries' income growth responds to changes in temperature and that estimates of β_1 and β_2 are highly robust to alternative specifications of the fixed effects and time controls[4].

An additional concern is that countries trade with one another and that unobserved temperature shocks across a trading network might lead to biased coefficient estimates in equation (1). However, if temperature shocks are uncorrelated across trading partners, then estimates of β_1 and β_2 still represent unbiased estimates of own-country temperature shocks on output; if shocks are correlated across trading partners, then β_1 and β_2 represent reduced-form estimates of the net effect in a given country of correlated shocks across that country's trading network. The main concern for our analysis is if the future pattern of temperature change should not correspond to the spatial pattern of historical shocks; however, we are unaware of any relevant research in climate science.

To quantify uncertainty in estimates of β_1 and β_2, we implement multiple bootstrapping strategies: (1) Sampling by country. From our list of 165 countries, draw (with replacement) a 165-element list of countries—which will omit some countries and contain duplicates of others—and retain all years of data for the selected countries; this is repeated 1,000 times, drawing a new country sample each time, re-estimating equation (1), and retaining estimates of β_1 and β_2. This approach allows for arbitrary correlation in residuals within countries over time. (2) Sampling by year. This allows for potential cross-sectional correlation in residuals in a given year, and is also repeated 1,000 times. (3) Sampling by five-year block. We divide the data into 10 five-year blocks (that is, 1961–65, 1966–70, and so on through 2010), and sample with replacement from these 10 blocks. This allows for both temporal and cross-sectional dependence in residuals, for example, as caused by global recessions that last multiple years.

Our main results use strategy (1) (sampling by country), but we show that our results are robust, regardless of the strategy used. In what follows, the bootstrapped response functions $h^j(T_{it}) = \hat{\beta}_1^{\,j} T_{it} + \hat{\beta}_2^{\,j} T_{it}^2$ are indexed with j, where $j \in (1, 2, \ldots, 1{,}000)$.

For each $h^j(T_{it})$, we define the 'temperature optimum' as the maximum of the quadratic function, that is, $\frac{-\beta_1^j}{2 \times \beta_2^j}$ (this is always a maximum because all estimates yield $\beta_1^{\,j} > 0$ and $\beta_2^{\,j} < 0$).

To ensure that equation (1) is capturing growth effects and not just level effects, we re-estimate equation (1) with additional lags of temperature (and their squares)[4,11]. This is important because countries' economic output could 'catch up' in the year following a temperature shock; this catch-up behaviour would not be captured in a model containing only contemporaneous temperature variables, but would be captured in a model that includes lags of temperature and where overall temperature effects are computed by summing contemporaneous and lagged coefficients[11]. We thus estimate equation (1) with up to five lags l of temperature, that is,:

$$h^j(T_{it}) = \sum_{l=0}^{5} \{\hat{\beta}_{1,l}^{\,j} T_{it-l} + \hat{\beta}_{2,l}^{\,j} T_{it-l}^2\} \quad (2)$$

and re-estimate all calculations below with results from these distributed lag models. Our main results with this sensitivity test are shown in Extended Data Fig. 4.

Climate model simulations. To follow the IPCC protocols, we analyse the exact climate model realizations and time periods used by the IPCC in its most recent assessment report[5]. These climate model realizations were generated by the World Climate Research Program under Phase Five of the Climate Model Intercomparison Project (CMIP5)[18]. For the historical baseline experiment, the CMIP5 protocol ran each climate model from the mid-1800s to 2005, using the historical climate forcings. For the future scenarios, the CMIP5 protocol used the RCPs, which assume different levels of climate forcing going forward in the 21st century. In total, there are four: RCP2.6, RCP4.5, RCP6.0 and RCP8.5.

Following the IPCC protocols, we use the same historical baseline period (1986–2005) and RCP future periods (2046–2065 and 2081–2100) as did the IPCC. In our bias correction method (see below), there are three RCPs whose global warming ranges are most consistent with the 1.5 °C and 2 °C targets in these IPCC scenario time periods: RCP4.5 and RCP6.0 during the 2046–2065 period, and RCP2.6 during the 2081–2100 period. (RCP2.6 is the only RCP scenario in which some models project global warming of less than 1.5 °C for the end of the century; for mid-century, none of the RCP2.6 model runs project warming above 2 °C, and so we do not utilize RCP2.6 for mid-century). We therefore calculate the distribution of GDP outcomes in response to the global warming levels projected during the 2046–2065 period of RCP4.5 and RCP6.0, and during the 2081–2100 period of RCP2.6. In addition, to compare the probability of economic impacts for the UN targets with the probability of those for higher levels of greenhouse gas emissions, we also calculate the distribution of GDP outcomes for the 2046–2065 and 2081–2100 periods of RCP8.5.

Uncertainty in the temperature-driven GDP impacts of a given level of greenhouse gas emissions arises from both uncertainty in the level of global warming associated with that level of emissions and uncertainty in the spatial pattern of temperature at that level of global warming. The IPCC climate analysis protocols span these uncertainty dimensions by analysing one realization of each climate model in each RCP scenario[5]. To follow the IPCC protocols, we analyse the same realizations as the IPCC.

However, it should be noted that the CMIP5 ensemble does not span the full range of each uncertainty dimension in a fully uniform framework. Rather, although the experimental conditions for the ensemble were coordinated between the modelling centres, both the models and the implementation of the simulation conditions vary across the ensemble. For example, the ensemble includes simulations from all national modelling centres that chose to participate, but not every modelling centre archived a simulation in each scenario. As a result, the IPCC selection of one realization of each model in each RCP yields different numbers of realizations—and model combinations—in each RCP (42 realizations in RCP4.5, 32 in RCP2.6, 25 in RCP6.0 and 39 in RCP8.5). Likewise, although each modelling centre conformed to a basic set of coordinated experimental conditions, the exact implementation of those conditions varied between the centres. This combination of coordinated but incomplete experimental uniformity has led the CMIP5 ensemble to be known as 'an ensemble of opportunity'. As in the IPCC, we leverage the CMIP5 ensemble of opportunity to estimate an approximate probability distribution; it should be emphasized that this approach is not identical to sampling across a probabilistic ensemble[27].

Because we use GDP data through 2010 and attempt to quantify economic impacts from that year forward, we must also project global and country-level temperature changes forward from the year 2010. To do so, and to control for individual climate model biases in average temperatures, we first calculate the difference between model-projected annual average future temperatures (in 2046–2065 or 2081–2100) and model-simulated annual average temperatures in the baseline 1986–2005 period. We then add those model-projected differences to the actual historical temperature observations.

For each climate model m corresponding to a chosen RCP scenario s at a given time period, we first calculate two quantities: (1). The magnitude of global temperature change ΔT^{sm}, which is the difference in annual average global surface temperature between a 1986–2005 baseline period and a future period (either 2046–2065 or 2081–2100). Gridded temperature projections relative to this baseline period are produced at 2.5° resolution. These are aggregated to a scalar 'global warming' projection by taking an average over all grid cells, with each cell g weighted by the cosine of the latitude of each cell g's centrepoint L (given the convergence of lines of latitude towards the poles):

$$\Delta T^{sm} = \frac{\sum_g \{\cos(\overline{L_g}) \times (T^{sm}_{g,\text{end}} - T^{sm}_{g,\text{base}})\}}{\sum_g \cos(\overline{L_g})} \quad (3)$$

(2). The magnitude of each country i's temperature change ΔT_i^{sm}, analogously computed by taking the average projected temperature change of all cells g but

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

weighted by their share of country *i*'s population P_{ig} rather than by their relative surface area. Gridded population distribution data[28] is provided at 30-arc-sec resolution and is aggregated to 2.5° resolution to match the temperature projection data. Thus, country-level temperature change projections are described by the equation:

$$\Delta T_i^{sm} = \frac{\sum_g \{P_{ig} \times (T_{g,\text{end}}^{sm} - T_{g,\text{base}}^{sm})\}}{\sum_g P_{ig}} \quad (4)$$

To express the future global-scale temperature values relative to pre-industrial values, as in the UN temperature targets, we add these model-projected differences between the future and the baseline to the global-scale warming that occurred between the pre-industrial period and the end of the period of GDP and temperature observations (which extends to 2010). According to the IPCC, the "globally averaged combined land and ocean surface temperature data as calculated by a linear trend, show a warming of 0.85 [0.65 to 1.06] °C, over the period 1880–2012," and the "total increase between the average of the 1850–1900 period and the 2003–2012 period is 0.78 [0.72 to 0.85] °C"[29]. We therefore assume that 0.8 °C of warming took place between the pre-industrial period and the end of our observational period. Thus for the global averages ΔT^{sm}, "global warming relative to pre-industrial" is equal to $\Delta T^{sm} + 0.8$ for all *s* and *m*.

To generate annual country-specific time series of projected future changes in temperature for input into the simulations below, we assume that temperatures increase linearly between the base period and the end period, and then add the linearized projected change in temperature to the observed average baseline temperature, thus 'bias-correcting' future national temperature time series. Thus for a given climate model–RCP realization, if the observed average historical temperature during the base period is $\overline{T_{i0}} = \frac{\sum_{t=1986}^{2005} T_{it}}{2005-1986}$, then the projected temperature in each future year is:

$$\Delta T_{it}^{sm} = \overline{T_{i0}} + \frac{t - t_{\text{base}}}{t_{\text{end}} - t_{\text{base}}} \times \Delta T_i^{sm} \quad (5)$$

where $t_{\text{base}} = 2010$ is the initial year of our simulation and t_{end} is either 2049 or 2099. (As before, small *t* indexes time and capital *T* refers to temperature). The assumed linear temperature increase appears to be consistent with RCP 4.5 or 6.0 through mid-century; it is perhaps less consistent with RCP2.6 through the end of the century, as RCP2.6 warms though mid-century and then stabilizes through to the end of the century. To understand whether our assumed linear warming path distorts our findings for RCP2.6, we conduct an additional experiment in which we assume all warming under RCP2.6 occurs by 2049, and then temperatures stabilize at this new level between 2050 and 2099 (Extended Data Fig. 5). This scenario has the same projected global warming by the end of the century as our baseline RCP2.6 scenario, but all warming is assumed to happen in the first half of the 21st century. As shown in Extended Data Fig. 5, we find that the scenario with rapid initial warming worsens the overall impacts of climate change and increases the cumulative benefits of limiting warming to 1.5 °C versus 2 °C.

Defining counterfactual growth scenarios. To project growth in GDP absent climate change, we use projections from the SSPs, a framework developed to describe conditions associated with various degrees of climate forcing by the end of the century. In all, there are five SSP narratives, each making different assumptions about mitigation and adaptation challenges, demographic trends, and developments in the energy industry[19]. We exploit the time series of projected country-level economic growth and population from 2010 to 2095 associated with the SSP1 narrative, because this appears to be the SSP most consistent with the forcing levels required to achieve 1.5 °C warming in 2049 or 2099[6] (although, as pointed out by ref. [6], with high enough carbon pricing all SSPs could potentially be consistent with 1.5 °C warming by mid-century, and three SSPs could be consistent with 1.5 °C warming by the end of the century). SSP1 is described as an optimistic future with 'low' challenges to adaptation and mitigation. SSP1 is characterized by many developing countries contributing an increasingly large share of global GDP by the end of the century (Extended Data Fig. 1a and b), with a larger share of total global GDP projected to be produced in countries with warmer average temperatures by the end of the century absent climate change (Extended Data Fig. 1c). In addition to using SSP1, we also test the robustness of our results to alternative choices from the other four SSPs (Fig. 4 and Extended Data Fig. 6).

Projecting economic impacts of 1.5 °C versus 2 °C. *Step (1). Assemble input data.* Required input data are the parameters of each response function $h^j(T_{it})$ estimated from each of the *j* bootstraps of equation (1); projections of country-year average temperature T_{it}^{sm} for each GCM *m* for a given RCP scenario *s* through to 2049 or 2099; projections of baseline country-year per capita growth rates λ_{it}^{κ} and populations ω_{it}^{κ} through 2099, for each country *i* and year *t*, from a given SSP scenario κ.
Step (2). Calculate country-specific growth trajectories for each bootstrap–RCP–GCM–SSP combination. Projections are initialized using average temperature and GDP per capita between 2000–2010 as the baseline for each of the countries in our analysis. For a given historical bootstrap run *j* and GCM–RCP–SSP projection $sm\kappa$, GDP per capita *y* in each future year $t+1$ in country *i* is projected by the equation:

$$y_{it+1}^{jsm\kappa} = y_{it}^{jsm\kappa} \times (1 + \lambda_{it+1}^{\kappa} + \varphi_{it+1}^{\kappa}) \quad (6)$$

where λ_{it+1}^{κ} is the level of economic growth projected by the data corresponding to the particular SSP series and $\varphi^{jsm} = h^j(T_{it+1}^{sm}) - h^j(T_{i0})$ is the additional estimated change in the growth rate due to the projected temperature increase above baseline for bootstrap run *j* and GCM projection *ms*. We also run a counterfactual no-warming scenario where temperatures are fixed at baseline levels, that is, $T_{it+1} = T_{i0}$ and $\phi_{it} = 0$ for all *i* and *t*):

$$y_{it+1}^{0,\kappa} = y_{it}^{0,\kappa} \times (1 + \lambda_{it+1}^{\kappa}) \quad (7)$$

With 165 countries, 1,000 bootstrap estimates of the temperature response function $h(\cdot)$, 100 total temperature time series (corresponding to 42, 25 and 32 climate models for mid-century RCP4.5, mid-century RCP6.0, and end-of-century RCP2.6, respectively, plus the constant-temperature series), five SSPs, and five bootstrap resampling schemes, we analysed more than 400 million distinct country-level economic pathways.
Step (3). Calculate global GDP trajectories for each bootstrap–RCP–GCM–SSP combination. For each GCM–bootstrap–SSP combination in a given period *t*, global GDP per capita is calculated as the average GDP per capita across countries, weighted by share of world population:

$$y_t^{jsm\kappa} = \sum_i \frac{\omega_{it}^{\kappa}}{\omega_t^{\kappa}} \times y_{it}^{jsm\kappa} \quad (8)$$

where $\frac{\omega_{it}^{\kappa}}{\omega_t^{\kappa}}$ is country *i*'s projected share of global population in year *t* for a given SSP. We similarly produce a time series of total global GDP by replacing $\frac{\omega_{it}^{\kappa}}{\omega_t^{\kappa}}$ with ω_{it}^{κ}, the country *i*'s projected population in that year. This is also calculated for the no-warming scenario, yielding counterfactual global GDP time series $y_t^{0,\kappa}$ and $Y_t^{0,\kappa}$, where Y_t denotes GDP.
Step (4). Calculate projected percentage changes in GDP or global GDP relative to the no-warming counterfactual for each bootstrap–RCP–GCM–SSP combination. For each bootstrap–RCP–GCM–SSP combination, we calculate the warming-induced percentage change in GDP relative to the counterfactual no-warming scenario in each country as:

$$\Psi_{it}^{jsm\kappa} = \frac{y_{it}^{jsm\kappa}}{y_{it}^{0,\kappa}} - 1 \quad (9)$$

This is calculated for $t = 2049$ for RCP4.5 and RCP6.0, and $t = 2099$ for RCP2.6. The percentage impact on global GDP per capita, $\Psi_t^{jsm\kappa}$, is calculated similarly for these endpoint years.
Step (5). Calculate projected discounted absolute changes in GDP or global GDP relative to the no-warming counterfactual for each bootstrap–RCP–GCM–SSP combination. The cumulative absolute dollar impact of warming is calculated for each country by taking the annual difference between the unique bootstrap–RCP–GCM–SSP projected GDP time series and the counterfactual no-warming time series, and discounting these differences back to present:

$$\Theta_i^{jsm\kappa} = \sum_t \frac{Y_{it}^{jsm\kappa} - Y_{it}^{0,\kappa}}{(1 + r_t)^{t - t_0}} \quad (10)$$

where $Y_{it}^{jsm\kappa} = y_{it}^{jsm\kappa} \times \omega_{it}^{\kappa}$ and r_t is the social discount rate that could vary with *t*. The global absolute impact is calculated by summing country-level impacts: $\Theta^{jsm\kappa} = \sum_i \Theta_i^{jsm\kappa}$.

Given the long-running and unresolved debate over how *r* should be specified, we calculate $\Theta_i^{jsm\kappa}$ under a range of approaches to specifying *r*. Specifically, we implement a variety of approaches discussed and implemented by previous authors, including implementations of the Ramsey equation with and without uncertainty and under alternate parameter choices for time preference and the marginal utility of consumption[30–34], calibrations to historical market interest rates in the USA[35,36], and constant discount rates[37] ranging from 2.5%–5%. Choices about the discount rate clearly have large implications for the estimation of damages. For instance, US$1,000 of damages in 50 years is worth US$228 today under a 3% annual discount rate, but only US$87 under a 5% annual rate.

As described by multiple authors[33,34,38], choices about *r* can be approached from the perspective of a social planner wishing to maximize the welfare of society. The central intuitions in this approach are that extra income or consumption is worth more to poor people than it is to rich people, and that with rising incomes a dollar

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

of additional income is worth less in the future than it is today. Under standard assumptions about the functional form of the 'utility function' that relates changes in consumption to changes in utility, this approach yields the Ramsey formula, which specifies the annual discount rate on consumption as:

$$r = \rho + \eta g \tag{11}$$

where ρ is the pure or social rate of time preference (the rate at which society discounts the utility of future generations), η is the elasticity of marginal utility of consumption (or how fast the utility of consumption declines as consumption increases), and g is the growth rate in consumption. If there is uncertainty about the growth rate in consumption, a third term is added to the Ramsey equation which induces a precautionary savings effect[34]:

$$r = \rho + \eta g - 0.5\eta^2\sigma_g^2 \tag{12}$$

where σ_g^2 is the variance in the growth rate. Uncertainty in future consumption growth enters negatively as the social planner, facing the possibility of slow future growth, wishes to transfer more resources to the future.

Using equations (11) and (12) and parameter choices about ρ and η from three benchmark studies[30–32] (Stern $\rho = 0.1, \eta = 1$; Nordhaus $\rho = 1, \eta = 2$; and Weitzman $\rho = 2, \eta = 2$; see Extended Data Fig. 1), we implement six versions of the Ramsey approach—three without uncertainty in future growth and three with uncertainty. For each bootstrap–RCP–GCM–SSP run, we define the growth rate g_t as the population-weighted average growth rate of GDP per capita:

$$g_t^{jsm\kappa} = \sum_i \frac{\omega_{it}^{\kappa}}{\omega_t^{\kappa}}(1 + \lambda_{it}^{\kappa} + \varphi_{it}^{jsm}) \tag{13}$$

with parameters defined as in equations (6) and (8) above. Average values across GCMs are shown in Extended Data Fig. 1a. Uncertainty in the growth rate for each future year is calculated as $\sigma_{gt}^2 = \mathrm{var}(g_t^{jsm\kappa})$, that is, the variance in projected growth rates in a given year across all bootstrap–RCP–GCM–SSP estimates. This probably represents a substantial lower bound on the true uncertainty in the growth rate, as it accounts only for uncertainty induced by additional warming and not for uncertainty in the underlying secular rate of growth (for which the SSPs do not provide uncertainty estimates).

Parameter choices and estimates of future growth rates are then used in either equation (11) or (12) to calculate year-specific discount rates r_t. The resulting estimates of Ramsey-based discount rates are shown in Extended Data Fig. 1b. All versions estimate higher interest rates in earlier periods, which is primarily a result of higher estimated baseline (SSP) growth rates in the earlier half of the century. Discount rates by end of century using the Ramsey approach range from 1.2% (Stern) to 4.2% (Weitzman), with the inclusion of the uncertainty term lowering discount rates only slightly.

Given that future baseline growth rates in developing and developed countries could be different, and given that the marginal effect of warming will probably differ between developing and developed countries given their different baseline temperatures, we also run scenarios where discount rates are allowed to differ between rich and poor countries (defined as being below or above the median level of GDP per capita at baseline). Specifically, using SSP1 data we produce separate population-weighted growth series for poor and rich countries (as shown in Extended Data Fig. 1c), and plug these growth projections into the Ramsey equation for each of the three benchmark choices of ρ and η to produce the six time series of discount rates that appear in Extended Data Fig. 1d. These income-specific discount rates, which are higher for poor countries than for rich countries given differences in baseline growth rates, are then applied to the relevant country groupings in the calculations below. As shown in Extended Data Fig. 3, allowing for income-specific discount rates results in higher median estimates of the global benefit of restricting warming to 1.5 °C. This is because global benefits are driven largely by impacts in the largest economies, including the USA and China, and allowing for income-specific discount rates lowers the rates for rich countries relative to the pooled scheme (for example, compare Extended Data Fig. 1b against Extended Data Fig. 1d), which translates to larger cumulative benefits in large economies projected to be harmed by warming (which again includes both the USA and China).

Beyond the Ramsey framework, another approach to specifying the discount rate uses the observed evolution of market interest rates over long periods combined with models of interest rate behaviour to project interest rates. We extract estimates from two of these exercises[35,36], both of which assume an initial interest rate of 4% and then project interest rates to fall by almost half by end of century (Groom and Newell-Pizer; Extended Data Fig. 1b). Unlike for the Ramsey discount rates, we assume these market discount rates are the same across bootstrap–RCP–GCM–SSP combinations, and just vary over time as shown in the plot.

For each bootstrap–RCP–GCM–SSP combination, each of these fourteen discount rates (six Ramsey with global average income, three Ramsey with rich/poor differences, two market-based, and fixed rates of 2.5%, 3% and 5%) are calculated for each and used in equation (10) to calculate the present value (in 2010) of the damages from warming.

Step (6). Calculate percentage or absolute damages at 1.5 °C versus 2 °C. To calculate relative damages at 1.5 °C versus 2 °C for a given bootstrap–RCP–SSP combination, we take estimates of percentage impacts $\Psi_{it}^{jsm\kappa}$ or discounted absolute impacts $\Theta_i^{jsm\kappa}$ across GCMs and fit a linear least-squares regression that relates estimated damages to the amount of global warming projected by the climate model by the end of the projection period (ΔT^{sm}). So for absolute damages in a given country, this regression is:

$$\Theta_i^{jsm\kappa} = \beta_i^{js\kappa}\Delta T^{sm} + \varepsilon_i \tag{14}$$

This relation is shown to be well approximated at the global level by a linear model (Fig. 1e–g). The slope of the linear fit $\beta_i^{js\kappa}$ is that bootstrap's estimate of the per-degree-Celsius impact of global temperature change on GDP per capita in country i. Halving this value thus gives us the impact of a half-degree change in global temperature for a given bootstrap, which, given linearity, is the estimated impact of limiting global warming to 1.5 °C relative to 2.0 °C in that country. Equation (14) is then re-estimated for each country and for each bootstrap, generating 1,000 estimates of impacts in each country for each RCP and SSP combination. We also estimate equation (14) at the global level to generate comparable results on percentage and absolute damages to global GDP. Global results are shown in Figs. 1 and 2, and country-level results are shown in Fig. 3a and b.

Step (7). Calculate probability of economic benefits of limiting warming to 1.5 °C versus 2 °C. Finally, we calculate the probability of economic gain under the 1.5 °C versus 2 °C scenarios—that is, the probability that damages from 1.5 °C of global warming will be smaller than damages from 2 °C of global warming—as the fraction of estimates of $\beta^{js\kappa}$ across 1,000 bootstrap runs that are negative. This is calculated for the world as a whole, as well as separately for each country (Fig. 3c and d).

Quantifying impacts of global warming beyond 2 °C. Recent estimates suggest that countries' current mitigation commitments (NDCs) are unlikely to limit global warming to 2 °C and are instead more likely to be consistent with warming in a 2.5–3 °C range[7]. To evaluate the impact of warming under these alternative warming outcomes, as well as for warming that exceeds 3 °C, we recalculate estimates of $\Psi_t^{jsm\kappa}$ and $\Theta^{jsm\kappa}$ across all RCPs s and for all SSPs κ. This provides estimates of the global impact of various warming scenarios relative to a no-warming counterfactual.

As shown in Fig. 4 and Extended Data Fig. 6, impacts are larger at higher levels of warming, with estimates suggesting that if current NDCs are achieved, global GDP could be 15%–25% lower by the end of the century as compared to a world that did not warm. Impacts for warming beyond 3 °C are even larger, but decline less steeply at the highest levels of warming (consistent with ref. [4]). This is because for hot countries that are substantially harmed by high levels of warming, GDP levels are bounded below by zero, whereas for cold countries that are substantially benefited by future warming, GDP levels can grow unbounded.

Quantifying sources of uncertainty in overall impacts of global warming. Our impact estimates (for example, on discounted global world product $\Theta^{jsm\kappa}$ from equation (10) above) are derived by combining historical regression results, future climate change projections from climate models, assumptions on baseline future growth rates from SSPs, and discount rates. Each of these has associated uncertainty, which we propagate throughout the analysis. In particular, total uncertainty in the impact of warming on global GDP under a given forcing scenario is a combination of uncertainty in how economies respond to warming (what we term 'historical regression uncertainty'), uncertainty across climate models in the amount and pattern of warming for a given level of forcing ('climate model uncertainty'), uncertainty in baseline future growth rates across SSPs ('SSP uncertainty'), and plausible alternatives for how to specify the discount rate ('discount rate uncertainty'). To quantify the relative contribution of each to overall impact uncertainty under a given level of forcing (RCP), we hold three out of four variables fixed and allow the fourth to vary. Variables are fixed as follows: historical regression uncertainty is fixed at the regression point estimate, discount rates are fixed at 3%, the SSP is fixed at the SSP providing the median impact estimate (typically SSP3), and the climate model projection is fixed at the model giving the median global warming projection for either mid-century or the end of the century.

Results for discounted cumulative global GDP loss due to warming are shown in Fig. 4b–d. For both 2049 (RCP4.5) and 2099 (RCP2.6), historical regression uncertainty—that is, uncertainty in how economies have responded to warming in the recent past—is the dominant source of uncertainty in overall impact projections for a given forcing level, followed by uncertainty due to alternative possible specifications of the discount rate. For instance, holding all other sources of uncertainty fixed for the end of the century, historical regression uncertainty alone leads to a 95% confidence interval of impact estimates of −US$122 trillion to

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

US$32 trillion, discount rate uncertainty to a 95% confidence interval of −US$375 trillion to −US$25 trillion, and climate model uncertainty to a 95% confidence interval of −US$78 trillion to US$4 trillion. Thus the overall uncertainty in impacts induced by uncertainty in economic parameters is around 2–4 times higher than that resulting from climate model uncertainty.

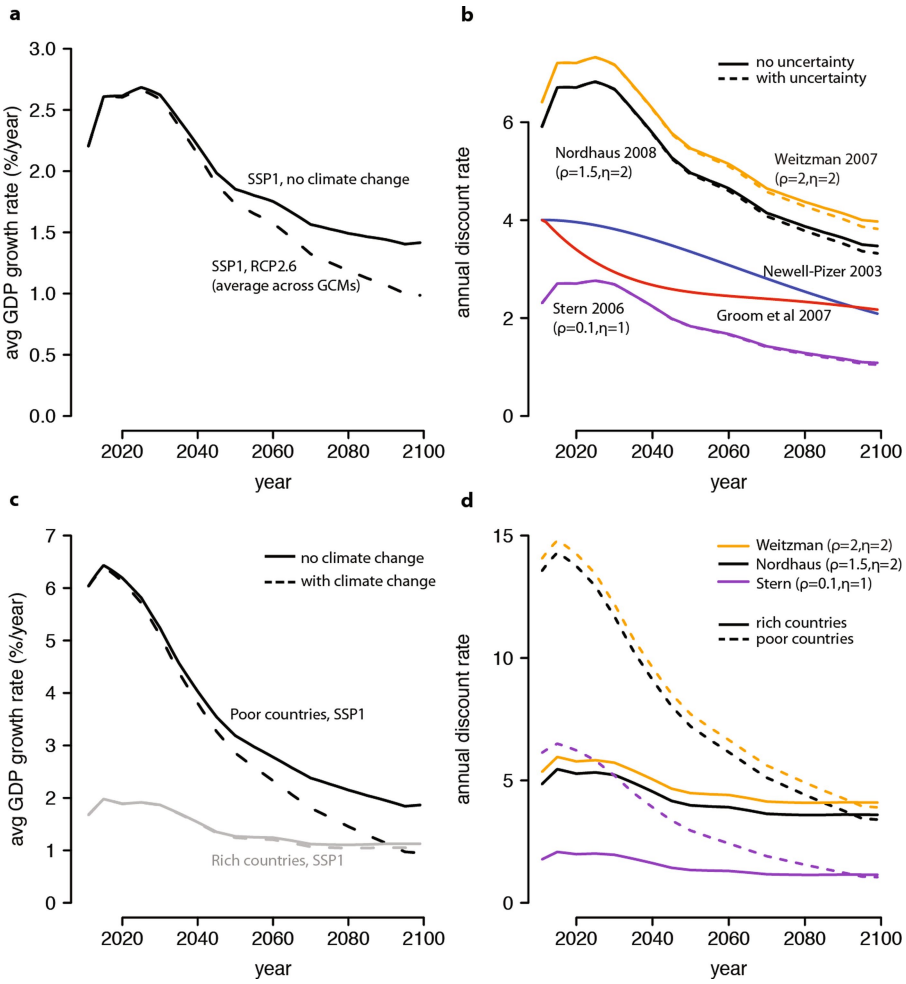
There are multiple caveats to this analysis, including that historical uncertainty would be larger if regression models with additional lags were also included, and that discount rate uncertainty could be understated if our 14 alternative discounting approaches do not span the range of 'plausible' discount rates.

While further constraining the range of plausible discount rates is perhaps challenging, not least owing to ethical considerations central to the choice of social-welfare-based discount rates[33], reducing uncertainty around how economies will respond to warming could be more tractable. Promising avenues could include detailed empirically based bottom-up assessments of climate impacts at the country level[23], leveraging existing sub-national or firm (company)-level data to estimate impacts[15,17], or using new fine-scale remote-sensing-based estimates of economic output to greatly increase the temporal and spatial specificity of outcome measurements[39,40].

Data availability. All data and code that support the findings of this study are available at https://purl.stanford.edu/vn535jm8926.

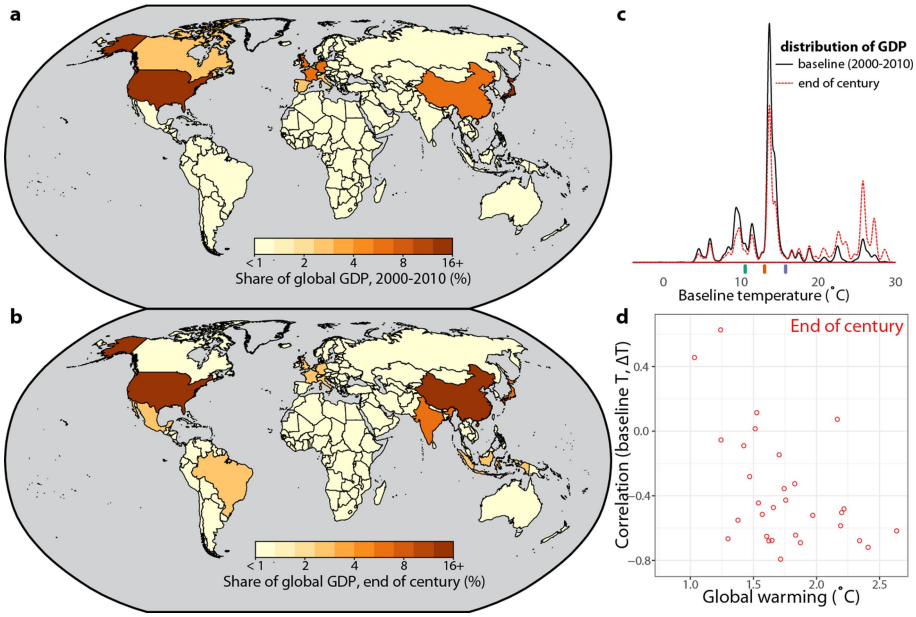
26. Matsuura, K. & Willmott, C. J. *Terrestrial Air Temperature: 1900–2010 Gridded Monthly Time Series* (version 3.01) (Center for Climatic Research, Univ. Delaware, Newark, 2012); http://climate.geog.udel.edu/~climate/.
27. Tebaldi, C. & Knutti, R. The use of the multi-model ensemble in probabilistic climate projections. *Phil. Trans. R. Soc. Lond. A* **365**, 2053–2075 (2007).
28. Balk, D. et al. Determining global population distribution: methods, applications and data. *Adv. Parasitol.* **62**, 119–156 (2006).
29. Hartmann, D. L. et al. Observations: atmosphere and surface. In Climate Change 2013: the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (eds Stocker, T. et al.) (Cambridge Univ. Press, Cambridge, 2013).
30. Stern, N. *The Economics of Climate Change: The Stern Review*. (Cambridge Univ. Press, Cambridge, 2006).
31. Weitzman, M. A review of the Stern review on the economics of climate change. *J. Econ. Lit.* **45**, 703–724 (2007).
32. Nordhaus, W. D. *A Question of Balance: Weighing the Options on Global Warming Policies*. (Yale Univ. Press, Yale, 2008).
33. Goulder, L. H. & Williams, R. C. III The choice of discount rate for climate change policy evaluation. *Clim. Change Econ.* **3**, 1250024 (2012).
34. Arrow, K. J. et al. Should governments use a declining discount rate in project analysis? *Rev. Environ. Econ. Policy* **8**, 145–163 (2014).
35. Newell, R. G. & Pizer, W. A. Discounting the distant future: how much do uncertain rates increase valuations? *J. Environ. Econ. Manage.* **46**, 52–71 (2003).
36. Groom, B., Koundouri, P., Panopoulou, E. & Pantelidis, T. Discounting the distant future: how much does model selection affect the certainty equivalent rate? *J. Appl. Econ.* **22**, 641–656 (2007).
37. Greenstone, M., Kopits, E. & Wolverton, A. Developing a social cost of carbon for us regulatory analysis: a methodology and interpretation. *Rev. Environ. Econ. Policy* **7**, 23–46 (2013).
38. Quiggin, J. Stern and his critics on discounting and climate change: an editorial essay. *Clim. Change* **89**, 195–205 (2008).
39. Henderson, J. V., Storeygard, A. & Weil, D. N. Measuring economic growth from outer space. *Am. Econ. Rev.* **102**, 994–1028 (2012).
40. Jean, N. et al. Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



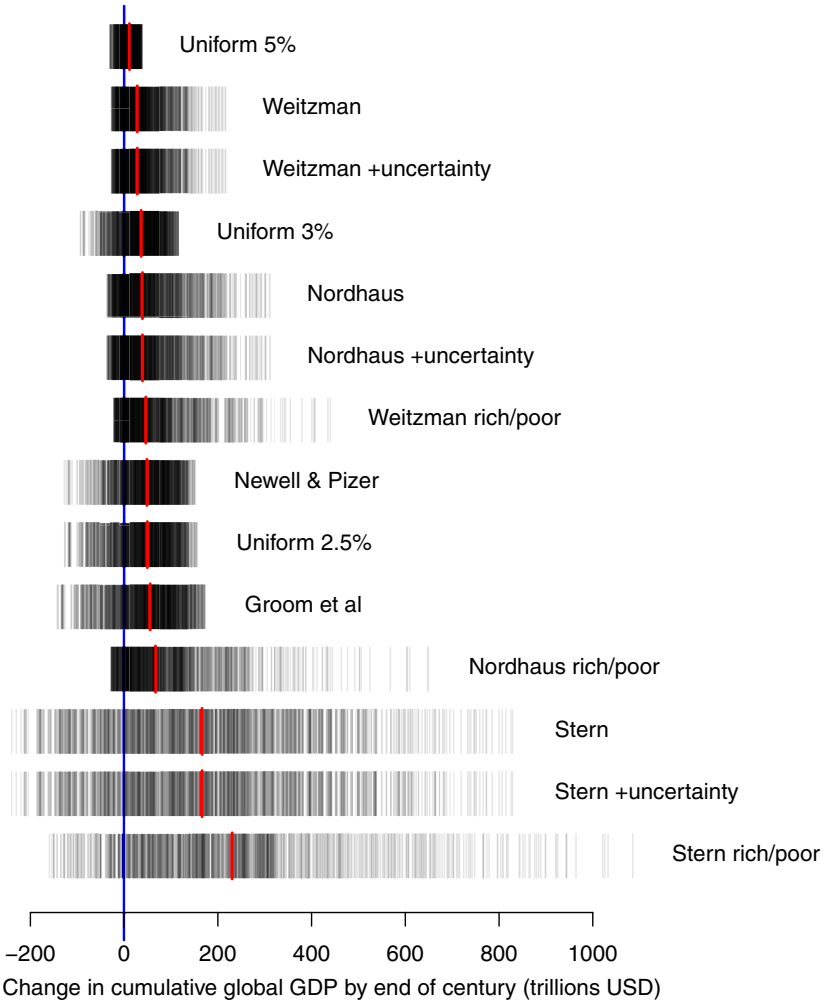
Extended Data Fig. 1 | Discount rate scenarios used in calculation of cumulative discounted impacts of future warming. a, Projected global average annual growth rates under SSP1 with and without climate change; estimates are averaged across bootstraps and climate models. Projected growth rates with climate change are used to define future consumption growth in Ramsey-based discount rates. **b**, Evolution of discount rates under different schemes through 2099. Ramsey-based schemes are Stern[30], Weitzman[31] and Nordhaus[32], with corresponding assumptions about the pure rate of time discount ρ and the elasticity of marginal utility of consumption η shown in parentheses. Dashed lines are versions of these Ramsey-based discounting schemes that account for growth-rate uncertainty. Non-Ramsey schemes are Newell and Pizer[35] and Groom[36]. **c**, Projected average annual growth rates separately for rich and poor countries under SSP1, with and without climate change. **d**, Corresponding Ramsey-based discount rates calculated separately for rich and poor countries, using income-specific growth rates from **c**.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



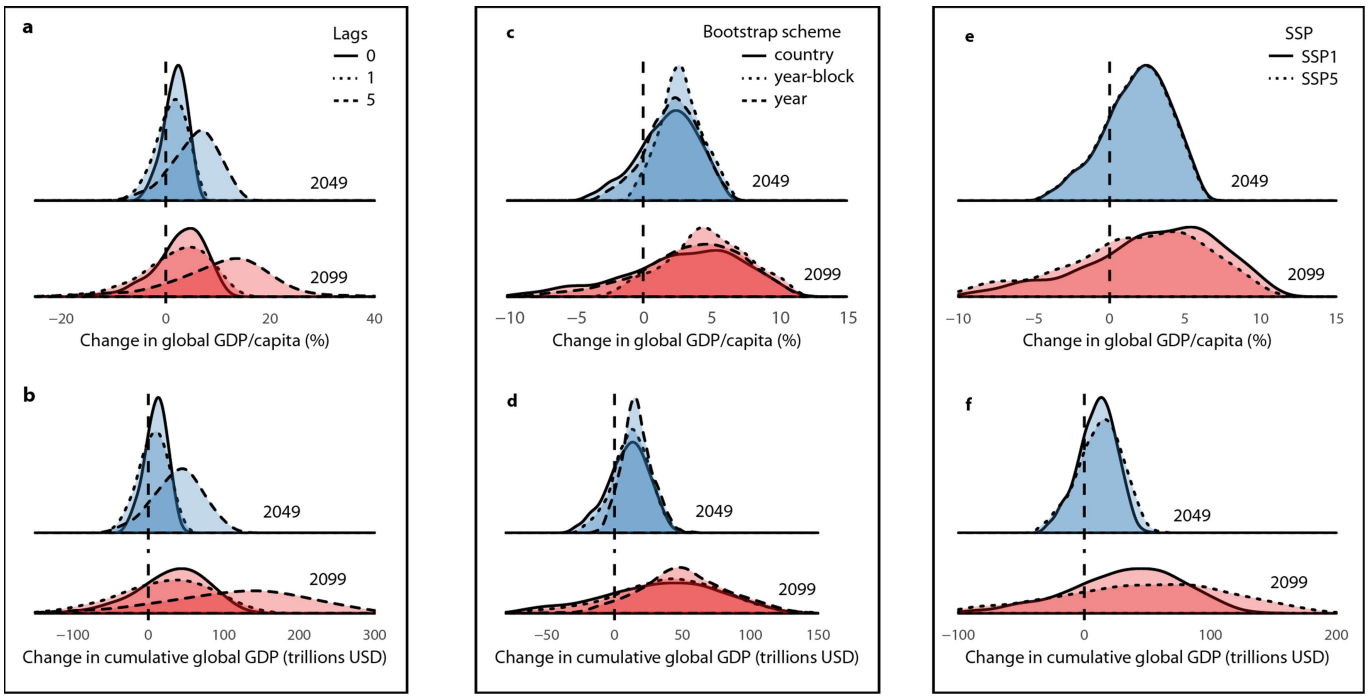
Extended Data Fig. 2 | Global GDP impacts can be negative at +1 °C but positive at +2 °C for some high-temperature-optimum bootstrap runs. a, **b**, Country share of global GDP at baseline (**a**) and by the end of the century (**b**) under SSP1, assuming no climate change. **c**, Distribution of global GDP by temperature, under baseline (black) and the end of the century SSP1 without climate change (red dashed); absent climate change, a substantial portion of global GDP is projected to be produced in countries with hotter average temperatures. **d**, Climate-model-predicted average global warming under RCP2.6 by the end of the century (*x* axis) versus the correlation between country-level baseline average temperature and country-level predicted warming in each model. In models that warm less at the global scale, countries that are currently warm tend to exhibit relatively larger warming, while in models that warm more at the global scale, countries that are currently cool tend to exhibit relatively larger warming. Future impacts on global GDP are a sum of country-specific impacts, which are a function of where each country is on the temperature response function (Fig. 1a) and the projected amount of future warming in that country; a given percentage impact in a country with a large GDP has a larger effect on global GDP than the same percentage impact in a country with small GDP. For high-temperature-optimum response functions (for example, Fig. 1g), impacts can be negative at +1 °C but positive at +2 °C because (i) absent climate change, a much larger proportion of total global GDP is projected by SSP1 to be produced in countries that are currently warmer than the optimum, and (ii) climate models with lower overall global warming projections under RCP2.6 tend to have higher relative warming in countries that are currently warm. This generates negative impacts at about 1 °C, where impacts are dominated by negative effects in warm countries (largely in the developing world), but positive impacts at about 2 °C, where high-latitude countries instead warm disproportionately and experience benefits that outweigh the damages in tropical countries.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



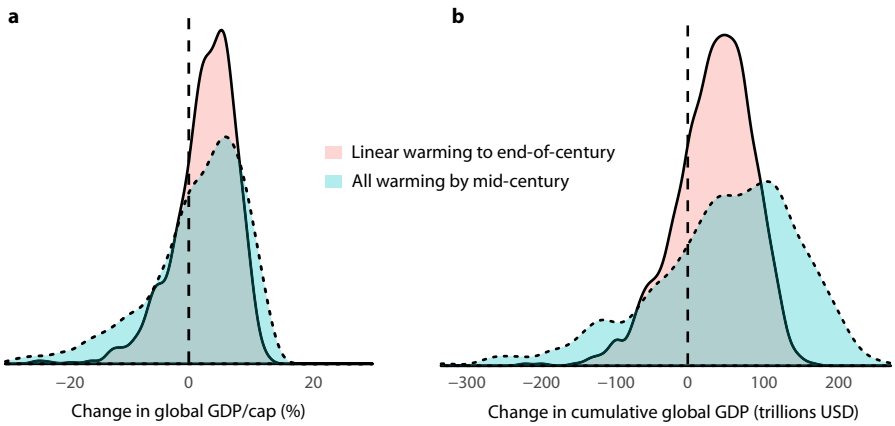
Extended Data Fig. 3 | Change in cumulative global GDP under 1.5 °C versus 2 °C global warming by the end of the century under different discounting schemes. Positive values indicate benefits (reduced losses) at 1.5 °C versus 2 °C. Each vertical line corresponds to a bootstrap estimate of benefits under each discounting scheme[30–32,35,36]. Red lines indicate median across bootstraps for each discounting scheme. Uniform schemes correspond to those in Extended Data Table 1; other schemes are described in Methods.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



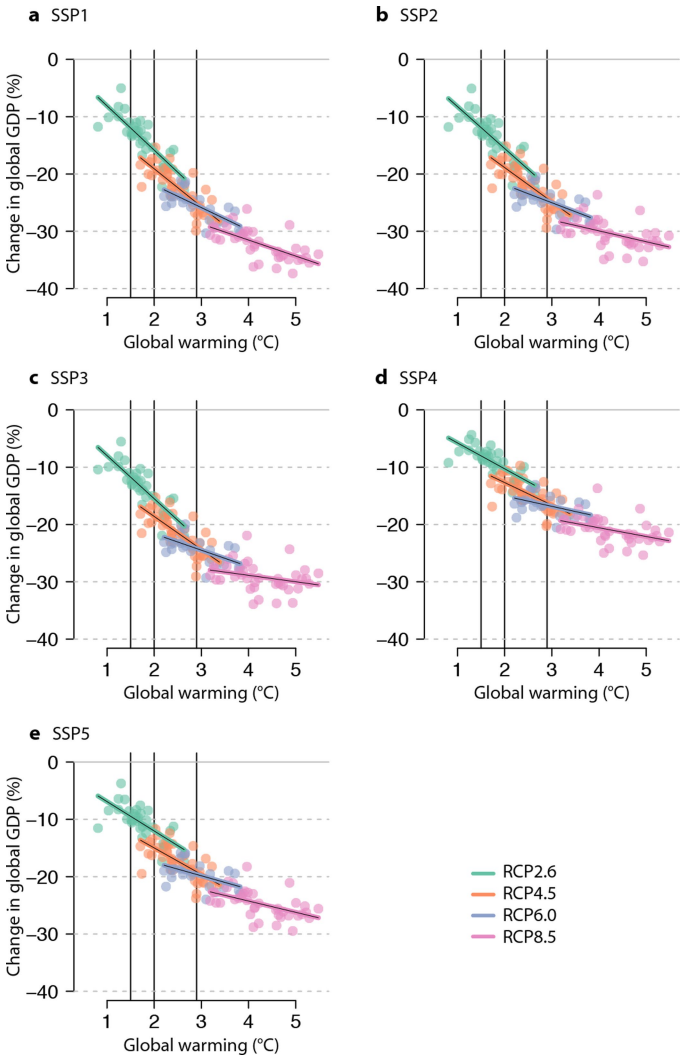
Extended Data Fig. 4 | Robustness of results to alternative specifications. Change in global GDP per capita in 2049 and 2099 based on regression models that include 0, 1 or 5 lags (**a** and **b**); bootstrap schemes that sample by country, five-year block or single year (**c** and **d**); or alternative SSPs (**e** and **f**). Top panels show percentage changes in global GDP per capita under 1.5 °C versus 2 °C; the bottom panels show change in cumulative global GDP in US$ trillions under a 3% discount rate.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 5 | Robustness under alternative warming paths. Benefit—in terms of per capita GDP (**a**) and cumulative GDP (**b**)—of 1.5 °C versus 2 °C by end of century under the baseline assumption that overall projected warming occurs linearly between the baseline year and 2099 (pink), versus projected benefit assuming that all projected warming occurs by 2049 and temperatures remain constant thereafter (blue). Both scenarios have the same projected global warming by the end of the century. For the same level of overall warming by the end of the century, scenarios with rapid initial warming worsen the overall impacts of climate change and increase the cumulative benefits of limiting warming to 1.5 °C versus 2 °C.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Projected change in global GDP (%) under global warming by the end of the century, for each SSP. Panels **a**–**e** show the change in GDP for different climate models under different RCP forcing scenarios, relative to a no-warming baseline (median bootstrap) for SSPs 1–5, respectively. Results are as in Fig. 4a, but for each SSP. Each dot represents an RCP-climate model projected change in global GDP under a given SSP; colours represent the four RCPs. Lines are least-squares fits to the points corresponding to the different RCPs with matching colour scheme. The three vertical black lines denote the 1.5 °C target, the 2 °C target and the median-estimated warming expected under current Paris commitments (2.9 °C)[7]. Warming is relative to pre-industrial levels.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Change in cumulative global GDP (in US$ trillions) under 1.5 °C versus 2 °C global warming by the end of the century under different discounting schemes

discount scheme	1%	5%	10%	25%	50%	75%	90%	95%	99%
Uniform 5%	-31	-18	-11	0	11	21	29	33	40
Weitzman	-29	-17	-12	4	28	65	112	144	217
Weitzman +uncertainty	-29	-17	-12	4	28	65	112	144	217
Uniform 3%	-95	-54	-32	3	36	65	88	101	119
Nordhaus	-39	-23	-16	5	39	91	160	206	312
Nordhaus +uncertainty	-39	-23	-16	5	39	91	160	206	313
Weitzman rich/poor	-22	-14	-6	11	46	96	176	244	442
Newell & Pizer	-130	-72	-43	5	49	86	115	132	156
Uniform 2.5%	-129	-73	-43	5	50	87	118	136	161
Groom et al	-145	-81	-48	5	56	98	132	151	179
Nordhaus rich/poor	-28	-17	-7	17	67	140	257	360	657
Stern	-240	-136	-88	24	166	336	515	618	840
Stern +uncertainty	-240	-136	-88	24	166	336	515	619	841
Stern rich/poor	-171	-99	-38	71	231	414	623	740	1091

Values show estimated impacts at different quantiles of the estimated impact distribution for each discounting scheme (uniform schemes[37], Weitzman[31], Nordhaus[32], Newell and Pizer[35], Groom[36] and Stern[30]), and correspond to estimates shown in Extended Data Fig. 3. Positive values indicate benefits (reduced losses) at 1.5 °C versus 2 °C.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Probability that limiting global warming to 1.5 °C will generate benefits relative to 2 °C warming

%Δglobal GDP/capita by mid-century		
benefit threshold	RCP4.5	RCP6.0
0%	0.80	0.73
1.25%	0.63	0.53
2.50%	0.42	0.31
3.75%	0.21	0.13
5.00%	0.07	0.03

Cumulative Δglobal GDP by mid-century						
	RCP4.5			RCP6.0		
benefit threshold	2.5%	3.0%	5.0%	2.5%	3.0%	5.0%
0	0.76	0.76	0.75	0.69	0.68	0.68
$10 trillion	0.56	0.53	0.32	0.47	0.43	0.23
$20 trillion	0.33	0.27	0.03	0.24	0.18	0.01
$30 trillion	0.14	0.08	0.00	0.08	0.04	0.00
$40 trillion	0.04	0.01	0.00	0.01	0.00	0.00

%Δglobal GDP/capita by end-of-century	
benefit threshold	RCP2.6
0%	0.76
2%	0.62
4%	0.45
6%	0.27
8%	0.12
10%	0.03

Cumulative Δglobal GDP by end-of-century			
	RCP2.6		
benefit threshold	2.5%	3.0%	5.0%
0	0.78	0.78	0.76
$10 trillion	0.72	0.71	0.53
$20 trillion	0.68	0.63	0.28
$50 trillion	0.50	0.38	0.00
$100 trillion	0.18	0.05	0.00
$150 trillion	0.02	0.00	0.00

Left panels show benefits in terms of percentage change in global GDP per capita by mid-century and the end of the century. For instance, by mid-century under RCP4.5 there is a 42% probability of benefits exceeding 2.5% of global GDP per cap. Right panels show benefits in terms of cumulative change in global GDP by mid-century and the end of the century, under three different discount rates for each relevant RCP. For instance, by the end of the century, there is a 50% probability of benefits exceeding US$50 trillion using a discount rate of 2.5%.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 3 | Probability that limiting global warming to 1.5 °C will generate different levels of benefits relative to 2.0 °C warming, under different discounting schemes

discount scheme	0	$10 trillion	$20 trillion	$40 trillion	$100 trillion	$200 trillion	$350 trillion
Uniform 5%	0.76	0.53	0.28	0.01	0.00	0.00	0.00
Weitzman	0.79	0.67	0.57	0.40	0.13	0.02	0.00
Weitzman +uncertainty	0.79	0.67	0.57	0.40	0.13	0.02	0.00
Uniform 3%	0.78	0.71	0.63	0.47	0.05	0.00	0.00
Nordhaus	0.80	0.70	0.63	0.50	0.22	0.06	0.01
Nordhaus +uncertainty	0.80	0.70	0.63	0.50	0.22	0.06	0.01
Weitzman rich/poor	0.86	0.76	0.68	0.54	0.24	0.08	0.02
Newell & Pizer	0.78	0.72	0.68	0.55	0.17	0.00	0.00
Uniform 2.5%	0.78	0.72	0.68	0.56	0.18	0.00	0.00
Groom et al	0.78	0.73	0.69	0.59	0.24	0.00	0.00
Nordhaus rich/poor	0.87	0.80	0.73	0.63	0.38	0.17	0.05
Stern	0.80	0.78	0.76	0.72	0.62	0.44	0.23
Stern +uncertainty	0.80	0.78	0.76	0.72	0.62	0.44	0.23
Stern rich/poor	0.86	0.85	0.83	0.80	0.70	0.54	0.32

Benefits are in terms of cumulative change in global GDP by the end of the century (RCP2.6). Discounting schemes are: uniform schemes[37], Weitzman[31], Nordhaus[32], Newell and Pizer[35], Groom[36] and Stern[30]).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0127-x

Inference of ecological and social drivers of human brain-size evolution

Mauricio González-Forero[1] & Andy Gardner[1]*

The human brain is unusually large. It has tripled in size from Australopithecines to modern humans[1] and has become almost six times larger than expected for a placental mammal of human size[2]. Brains incur high metabolic costs[3] and accordingly a long-standing question is why the large human brain has evolved[4]. The leading hypotheses propose benefits of improved cognition for overcoming ecological[5–7], social[8–10] or cultural[11–14] challenges. However, these hypotheses are typically assessed using correlative analyses, and establishing causes for brain-size evolution remains difficult[15,16]. Here we introduce a metabolic approach that enables causal assessment of social hypotheses for brain-size evolution. Our approach yields quantitative predictions for brain and body size from formalized social hypotheses given empirical estimates of the metabolic costs of the brain. Our model predicts the evolution of adult *Homo sapiens*-sized brains and bodies when individuals face a combination of 60% ecological, 30% cooperative and 10% between-group competitive challenges, and suggests that between-individual competition has been unimportant for driving human brain-size evolution. Moreover, our model indicates that brain expansion in *Homo* was driven by ecological rather than social challenges, and was perhaps strongly promoted by culture. Our metabolic approach thus enables causal assessments that refine, refute and unify hypotheses of brain-size evolution.

The leading hypotheses for the evolution of brain size make different suggestions as to which cognitive challenges have been the most important in driving brain expansion. 'Ecological-intelligence' hypotheses emphasize challenges posed by the non-social environment, for example, finding, caching or processing food[5–7] (Fig. 1). By contrast, 'social-intelligence' hypotheses emphasize challenges posed by the social environment, for example, cooperating for resource extraction[10,15], manipulating others, avoiding manipulation or forming coalitions and alliances to outcompete others[8,9] (Fig. 1). Social challenges have been suggested to constitute particularly powerful drivers of brain expansion, because they may have triggered evolutionary arms races in cognition[8–10]. Finally, 'cultural-intelligence' hypotheses emphasize challenges of learning from others, teaching and doing so when there is accumulated cultural knowledge[11–14]. Empirical tests of these hypotheses customarily investigate phylogenetically controlled correlations between brain size (or the size of brain components) and candidate selective factors (for example, diet type[5,17], tactical-deception rate[18], group size[10,19] and social-learning proclivity[20]). However, establishing causality has proven to be difficult. For example, given a positive correlation, it is unclear whether large group sizes favour bigger brains or big brains enable larger group sizes[16]. Moreover, there is the quantitative problem of explaining not only why bigger brains are favoured, but also why they are favoured to the particularly large size observed in humans (around 1.3 kg for a body size of approximately 50 kg in females[21,22]).

To address these problems, we merge elements of metabolic theory[23], life-history theory and differential games to obtain quantitative predictions for the evolution of brain and body size when individuals face ecological and social challenges given metabolic costs of the brain. Our approach incorporates social interactions into a previous non-social model[24] (Supplementary Information 1–3). As a first approximation, we consider a female population and partition the body mass of each individual into three tissues: 'brain', 'reproductive' and other 'somatic' tissue (Fig. 2a). Part of the energy consumption of reproductive tissue is for the production and maintenance of offspring, whereas part of energy consumption of the brain is for production (learning) and maintenance (memory) of energy-extraction skills. Accordingly, at each age the individual has a certain skill level measured in information units (that is, bytes). She extracts energy by using her skill level to overcome ecological or social energy-extraction challenges. Success in an ecological challenge depends on her own skill level, whereas success in a social challenge depends on her skill level and that of her social partners. We

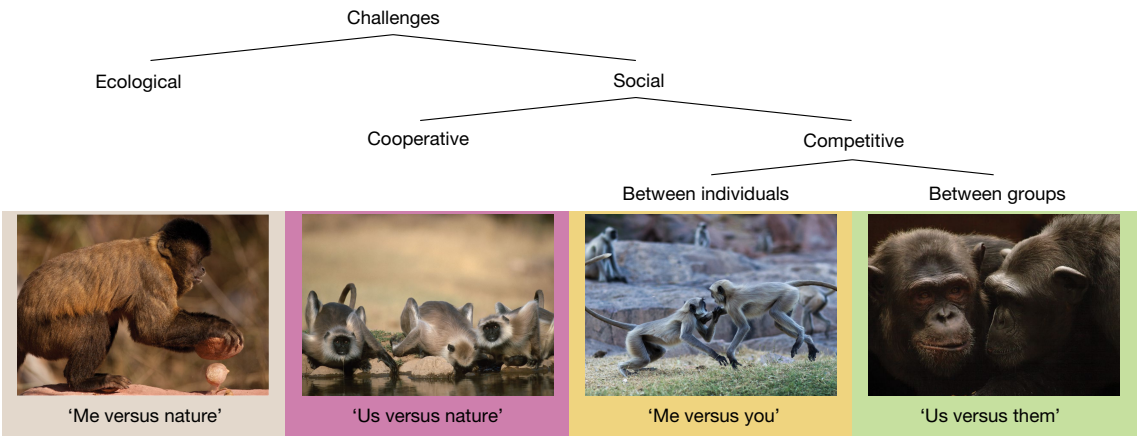


Fig. 1 | Ecological and social hypotheses for brain expansion. Ecological hypotheses emphasize challenges 'against nature', whereas social hypotheses emphasize challenges involving social partners. Here we partition these hypotheses into four types of challenges that are expected to trigger different evolutionary processes.

[1]School of Biology, University of St Andrews, St Andrews, UK. *e-mail: mgf3@st-andrews.ac.uk

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

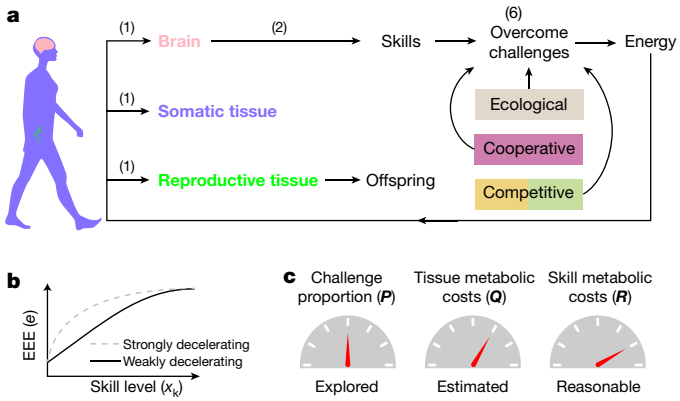


Fig. 2 | Model description. a, Schematic description of the model (see text for details). Numbers in parentheses indicate the corresponding equations in the Methods. **b**, **c**, The model depends on the shape of EEE with respect to skill (**b**) and three sets of parameters $\boldsymbol{P}$, $\boldsymbol{Q}$ and $\boldsymbol{R}$ (**c**).

consider three types of social challenge: 'cooperative', in which the individual's skill level and that of a social partner of the same age (hereafter 'peer') interact to overcome a challenge; 'between-individual competitive', in which the individual uses her skill level against that of a peer to extract energy; and 'between-group competitive', in which the individual's skill level and that of a peer act together and against the skills of another two peers (that is, one coalition competing against another). We assume that during any small time-interval, the individual faces energy-extraction challenges, a proportion (P_j) of which are of type j (collectively denoted by $\boldsymbol{P}$, with $j = 1,\ldots, 4$ indexing the four challenge types and $\sum_{j=1}^{4} P_j = 1$). Given $\boldsymbol{P}$, the growth strategy controls the amount of energy allocated to the production of each tissue throughout life, and we let the growth strategy evolve. The individual's energy-extraction efficiency (EEE) thus depends on her skill level, the skill levels of her social partners and the challenges faced, and we consider two shapes for EEE (Fig. 2b and Extended Data Fig. 1). We use previously published data, primarily from Kuzawa et al.[21], for parameter estimates, including metabolic costs of the brain (Fig. 2c, Methods and Extended Data Figs. 2, 3).

We find that increasing the proportion of cooperative challenges decreases both adult absolute brain size (hereafter 'brain size') and adult relative brain size (hereafter 'encephalization quotient', which is the adult brain size divided by expected brain size for a given body size[2]; Fig. 3a–c and Extended Data Fig. 4). By contrast, increasing the proportion of between-individual competitive challenges increases brain size when EEE is weakly decelerating with skill (Fig. 3a), but decreases brain size when EEE is strongly decelerating (Fig. 3b and Extended Data Fig. 4). However, although between-individual competition increases brain size with weakly decelerating EEE, the result is larger brains and smaller bodies than those of modern humans (Fig. 3a and Extended Data Fig. 4). Between-individual competition also decreases body mass as it increases the difficulty of energy extraction and thus limits the energy available for body growth; consequently, between-individual competition increases the encephalization quotient, either because brain size increases and body size decreases or because brain size decreases and body size decreases more strongly than brain size. Increasing the proportion of between-group competition generally decreases brain size, but increases the encephalization quotient, because body size decreases more strongly than brain size (Fig. 3a, b). However,

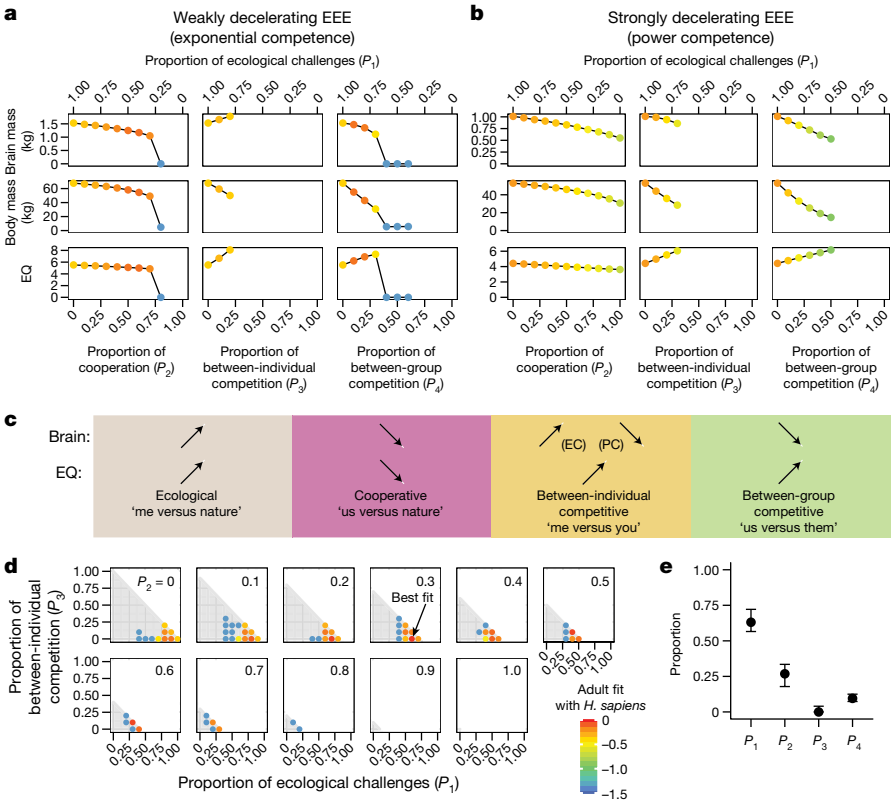


Fig. 3 | Effects of challenge types on brain size and best-fitting scenario for adult *H. sapiens*. a, **b**, Effects of increasing the proportion of a challenge type while decreasing the proportion of ecological challenges. EQ, encephalization quotient. **a**, Weakly decelerating EEE (exponential competence). **b**, Strongly decelerating EEE (power competence). Dot colour indicates the adult fit with brain and body mass of *H. sapiens* (that is, $-D(\tau_a)$; Supplementary Information 6). Zero adult fit means a perfect fit. **c**, Qualitative effects of challenge types on brain mass and encephalization quotient. **d**, Best fitting scenario for *H. sapiens*. Dots indicate the adult fit for every challenge combination that was solvable. Shaded regions indicate the simplex in which P_j can vary. The best fit occurs in $\boldsymbol{P}^* = (0.6, 0.3, 0, 0.1)$ (adult fit: −0.03). **e**, 'High fit' intervals around $\boldsymbol{P}^*$ where adult fit is greater than −0.05 and dots are interpolated points with the best fit.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

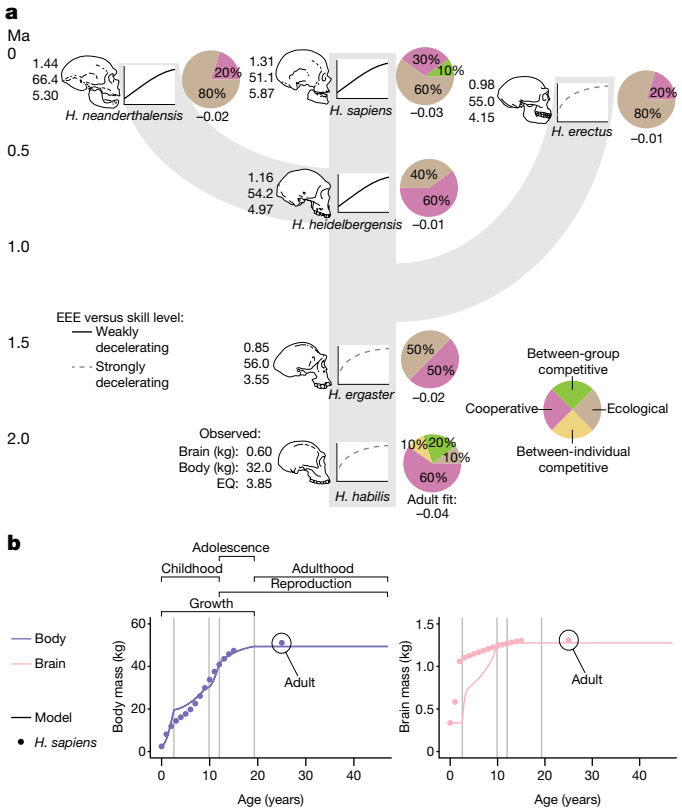


Fig. 4 | Best-fitting scenarios across adult *Homo*, and resulting predicted life history for *H. sapiens*. **a**, Best adult fitting scenarios across *Homo* (figure modified with permission from figure 8.1 of ref. [1]). Pie charts and plots respectively show the challenge combination and shape of EEE versus skill that yielded the best adult fit using the same $\boldsymbol{Q}$ and $\boldsymbol{R}$ parameters (Extended Data Figs. 6, 7). **b**, Life history with the challenge combination yielding the best adult fit for *H. sapiens*. Resulting life periods are indicated above the body mass plot. Vertical lines are ages at which the growth strategy changes suddenly; within childhood, they occur when brain growth begins and terminates.

with weakly decelerating EEE and additive cooperation, between-group competition increases both brain size and the encephalization quotient (Extended Data Fig. 4). Moderately frequent between-individual or between-group competition can lead to no allocation to brain and body growth (blue dots in Fig. 3a and Extended Data Fig. 4; see also Extended Data Fig. 5a, d); additionally, moderately frequent between-group competition in the presence of substantial cooperation can lead to arms races in brain size, which fail to yield stable, large brains (for example, because of cycling in brain size or eventual collapse to no allocation into brain growth (Extended Data Fig. 5)). This is because energy extraction becomes exceedingly difficult in the presence of large-brained competitors such that investments in brain or body growth do not pay off and the individual instead invests in early reproduction.

To determine if any combination of social-challenge parameters $\boldsymbol{P}$ yields an accurate prediction of adult brain and body sizes of *H. sapiens* and closely related species, we obtained solutions exhaustively across the $\boldsymbol{P}$ parameter space while holding the other parameters ($\boldsymbol{Q}$ and $\boldsymbol{R}$) fixed (Fig. 3d, e; Supplementary Information 5). We find near-perfect adult fits across *Homo* species (Fig. 4a and Extended Data Figs. 6–8). A near-perfect adult fit for *H. sapiens* occurs with a large proportion of ecological challenges (approximately 60%), a moderate proportion of cooperative challenges (around 30%), a small proportion of between-group competitive challenges (around 10%), and an approximately complete absence of between-individual competitive challenges (around 0%) (Figs. 3e, 4a and Extended Data Figs. 6, 9). In the resulting reconstruction for *Homo*, ecological challenges increase brain size whereas social challenges decrease it (Extended Data Fig. 4), the proportion of ecological challenges tends to increase from early to late *Homo* species, and a steep increase in encephalization quotient from *Homo ergaster* to *Homo heidelbergensis* is due to a transition from strongly to weakly decelerating EEE (Fig. 4a). The adult best-fit eco-social scenario for *H. sapiens* also yields a predicted life history that closely matches the species' life-history timing (Fig. 4b and Extended Data Fig. 10). The resulting ontogenetic fit is high for body size, but lower for brain size early in ontogeny (Fig. 4b), perhaps caused in part by our use of a power-law relationship between resting metabolic rate and body mass that underestimates resting metabolic rate early in the ontogeny[24]. With the adult brain size resulting from the best-fitting scenario for *H. sapiens* ($x_b^*(t_a) = 1.276$ kg) , the predicted adult skill level for energy extraction is 3.92 terabytes (TB), which can be calculated with an equation transforming brain mass to skill level[24] ($\hat{x}_k = s_k B_b x_b^*(t_a)/B_k$ where $\hat{x}_k$ is the asymptotic skill level in adulthood; equation (5) in the Methods). By comparison, current rough estimates[25] suggest a human-neocortex storage capacity of approximately 600 TB (Supplementary Information 4.3).

Using previously published data[21] for parameter estimates, our results suggest that adult human-sized brains and bodies may result from ecological challenges as drivers of brain expansion, with cooperation and between-group competition decreasing brain and body size and between-group competition increasing the encephalization quotient by decreasing body size more strongly than brain size (Fig. 3a and Extended Data Fig. 4b). In this eco-social scenario, between-individual competition is unimportant, as it does not lead to human-sized brains and bodies. Cooperation decreases brain size, because it allows individuals to rely on their partners' skills and thus decrease their own investment into costly brains (cooperation invites cheating), which is consistent with diminished brain sizes in cooperatively breeding birds[26] and mammals[27], including primates[28]. For instance, among mole rats, naked mole rats are the most specialized in cooperative breeding and have the smallest relative brain size[29] (however, allomaternal care and brain size are positively associated in mammals[30], but allomaternal care constitutes cooperation targeted at young, which vanishes in adulthood as opposed to the peer-cooperation studied here). Similarly, between-group competition can decrease brain size probably because between-group competition involves cooperation between group members, allowing individuals to rely on their partners' skill. The result that exceedingly frequent competition decreases absolute and relative brain size may be relevant to the observed decreased brain size in cetaceans with the largest group sizes[19]. Cooperation can also decrease body size in our model, because when brain size is disfavoured so too can be body size. This is because a consequence of our model is that a key reason to grow somatic tissue is to make energy available for brain growth: increasing the mass of inexpensive somatic tissue can increase the energy available for tissue (and brain) growth due to the physical constraint imposed by the power-law relationship between resting metabolic rate and body mass[24].

Overall, our assessment fails to support social hypotheses as explanations for the evolution of human brain size, and is more consistent with ecological hypotheses. Our results suggest causal interpretations that differ from some current thinking on the evolution of human cognition. Specifically, we obtained an eco-social scenario that involves a substantial proportion of cooperation (30% against nature and 10% against others), which could shape cognition towards cooperation. This would help to explain aspects of human cognition that facilitate cooperation[11], even if cooperation has not been a driver of human brain expansion. Additionally, because our analysis suggests that brain expansion in *Homo* has not been driven by peer cooperation or competition, our results indicate that social complexity may have had a more limited role in human brain size expansion than is commonly thought—although we emphasize that our analysis is an illustrative starting point and future extensions are encouraged (see Supplementary Information 9). Therefore, our results highlight the fundamental question of why ecological challenges would have favoured substantial brain

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

expansion in humans but less so in other taxa[10,15]. One clue is suggested by our finding that *H. sapiens*-sized brains and bodies can be obtained only under weakly decelerating EEE (Figs. 3, 4 and Extended Data Fig. 6a): in other words, only when young individuals can maintain a substantial rate of increase in their efficiency of energy extraction as they acquire skills. One possibility is that culture (or cumulative culture) facilitates weakly decelerating EEE if learning from the pool of skills in the population allows individuals to maintain a relatively high rate of increase in EEE as their skill level increases when young. More specifically, the evolution of progressively elaborated social learning, teaching and language[11–14] may have enabled young individuals to continue gaining skills with age, possibly promoting less strongly decelerating EEE. In this respect, our results are consistent with aspects of various cultural hypotheses for brain evolution[13,14] and an explicit account of the effect of culture on EEE could help to address whether culture (or cumulative culture) has enabled ecological challenges to drive brain expansion in humans in ways that have not occurred in other taxa.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0127-x.

Received: 6 December 2017; Accepted: 26 March 2018;
Published online 23 May 2018.

1. Klein, R. G. *The Human Career*. 3rd edn, (Univ. Chicago Press, Chicago, 2009).
2. Martin, R. D. Relative brain size and basal metabolic rate in terrestrial vertebrates. *Nature* **293**, 57–60 (1981).
3. Aiello, L. C. & Wheeler, P. The expensive-tissue hypothesis: the brain and the digestive system in human and primate evolution. *Curr. Anthropol.* **36**, 199–221 (1995).
4. Darwin, C. *The Descent of Man, and Selection in Relation to Sex*. (J. Murray, London, 1871).
5. Clutton-Brock, T. H. & Harvey, P. H. Primates, brains and ecology. *J. Zool.* **190**, 309–323 (1980).
6. Gibson, K. R. in *Primate Ontogeny, Cognition and Social Behaviour* (eds Else, J. G. & Lee, P. C.) 93–103 (Cambridge Univ. Press, Cambridge, 1986).
7. Rosati, A. G. Foraging cognition: reviving the ecological intelligence hypothesis. *Trends Cogn. Sci.* **21**, 691–702 (2017).
8. Humphrey, N. K. in *Growing Points in Ethology* (eds Bateson, P. P. G. & Hinde, R. A.) 303–317 (Cambridge Univ. Press, Cambridge, 1976).
9. Byrne, R. & Whiten, A. (eds). *Machiavellian Intelligence*. (Oxford Univ. Press, Oxford, 1988).
10. Dunbar, R. I. M. The social brain hypothesis. *Evol. Anthropol.* **6**, 178–190 (1998).
11. Moll, H. & Tomasello, M. Cooperation and human cognition: the Vygotskian intelligence hypothesis. *Phil. Trans. R. Soc. B* **362**, 639–648 (2007).
12. van Schaik, C. P. & Burkart, J. M. Social learning and evolution: the cultural intelligence hypothesis. *Phil. Trans. R. Soc. B* **366**, 1008–1016 (2011).
13. Henrich, J. *The Secret of our Success*. (Princeton Univ. Press, Princeton, 2016).
14. Laland, K. N. *Darwin's Unfinished Symphony*. (Princeton Univ. Press, Princeton, 2017).
15. Dunbar, R. I. M. & Shultz, S. Why are there so many explanations for primate brain evolution? *Phil. Trans. R. Soc. B* **372**, 20160244 (2017).
16. Powell, L. E., Isler, K. & Barton, R. A. Re-evaluating the link between brain size and behavioural ecology in primates. *Proc. R. Soc. B* **284**, 20171765 (2017).
17. DeCasien, A. R., Williams, S. A. & Higham, J. P. Primate brain size is predicted by diet but not sociality. *Nat. Ecol. Evol.* **1**, 0112 (2017).
18. Byrne, R. W. & Corp, N. Neocortex size predicts deception rate in primates. *Proc. R. Soc. B* **271**, 1693–1699 (2004).
19. Fox, K. C. R., Muthukrishna, M. & Shultz, S. The social and cultural roots of whale and dolphin brains. *Nat. Ecol. Evol.* **1**, 1699–1705 (2017).
20. Street, S. E., Navarrete, A. F., Reader, S. M. & Laland, K. N. Coevolution of cultural intelligence, extended life history, sociality, and brain size in primates. *Proc. Natl Acad. Sci. USA* **114**, 7908–7914 (2017).
21. Kuzawa, C. W. et al. Metabolic costs and evolutionary implications of human brain development. *Proc. Natl Acad. Sci. USA* **111**, 13010–13015 (2014).
22. Dekaban, A. S. & Sadowsky, D. Changes in brain weights during the span of human life: relation of brain weights to body heights and body weights. *Ann. Neurol.* **4**, 345–356 (1978).
23. West, G. B., Brown, J. H. & Enquist, B. J. A general model for ontogenetic growth. *Nature* **413**, 628–631 (2001).
24. González-Forero, M., Faulwasser, T. & Lehmann, L. A model for brain life history evolution. *PLoS Comput. Biol.* **13**, e1005380 (2017).
25. Bartol, T. M. Jr et al. Nanoconnectomic upper bound on the variability of synaptic plasticity. *eLife* **4**, e10778 (2015).
26. Fedorova, N., Evans, C. L. & Byrne, R. W. Living in stable social groups is associated with reduced brain size in woodpeckers (*Picidae*). *Biol. Lett.* **13**, 20170008 (2017).
27. Clutton-Brock, T. *Mammal Societies*. (Wiley, Oxford, 2016).
28. Thornton, A. & McAuliffe, K. Cognitive consequences of cooperative breeding? A critical appraisal. *J. Zool. (Lond.)* **295**, 12–22 (2015).
29. Kruska, D. C. T. & Steffen, K. Encephalization of Bathyergidae and comparison of brain structure volumes between the Zambian mole-rat *Fukomys anselli* and the giant mole-rat *Fukomys mechowii*. *Mamm. Biol.* **74**, 298–307 (2009).
30. Isler, K. & van Schaik, C. P. Allomaternal care, life history and brain size evolution in mammals. *J. Hum. Evol.* **63**, 52–63 (2012).

Acknowledgements We thank L. Lehmann for advice on the model set-up and for support at the University of Lausanne, where this project was initiated; P. Avila, S. H. Montgomery, S. B. Laughlin, M. Falcone, T. Clutton-Brock, and M. B. Morrissey for discussions; and A. Whiten for feedback. M.G.-F. was funded by a Marie Skłodowska-Curie Individual Fellowship (No 701464) and A.G. was funded by a NERC Independent Research Fellowship (NE/K009524/1). The contents of this paper reflect only the authors' view and not the views of their funders.

Reviewer information *Nature* thanks A. R. DeCasien and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.G.-F. conceived and carried out the study. A.G. conceived implementing challenge proportions during small time intervals and how to obtain high fit intervals. M.G.-F. and A.G. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0127-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0127-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.G-F.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Here we summarize our model; see Supplementary Information for a detailed description of the methods. No experiments were performed in this study and no empirical data were collected.

Model description. We consider a female population with overlapping generations and partition the body mass of each individual into three tissues: brain, reproductive and other somatic tissues. Each individual extracts energy from the environment at each time point to grow and maintain her tissues. We assume that some of the energy consumed by reproductive tissue is for production and maintenance of offspring, while some of that consumed by brain tissue is for production (learning) and maintenance (memory) of domain-general, energy-extraction skills. Accordingly, at each age the individual has a certain skill level measured in information units (that is, bytes). She extracts energy by using her skill level to overcome energy-extraction challenges that can be ecological or social. Success in an ecological challenge depends on her skill level and on the challenge difficulty, which is determined by the (non-evolving) environment. By contrast, success in a social challenge depends on her skill level and that of her social partners. We consider three different types of social challenge: cooperative challenges, in which the individual's skill level and that of a social partner of the same age (peer) interact to overcome a challenge that has a difficulty that is determined by the environment; between-individual competitive challenges, in which the individual uses her skill level against that of a peer to extract energy, such that the difficulty of energy extraction is determined by her competitor's skill level; and between-group competitive challenges, in which the individual's skill level and that of a peer act together and against the skills of another two peers (that is, one coalition competing against another), which determines the challenge difficulty. During any small time-interval, the individual faces challenges, a proportion P_j of which are of type j (with $j = 1, \ldots, 4$ indexing the four challenge types and $\sum_{j=1}^{4} P_j = 1$). For instance, $P_1 = 1$ denotes that individuals face only ecological challenges, whereas $P_1 = P_2 = 0.5$ denotes that individuals only face ecological and cooperative challenges and with equal proportions. We define the growth-metabolic rate as the rate of heat release by a resting individual due to tissue production. Moreover, we define the growth strategy as the fraction of the growth-metabolic rate due to the production of each tissue throughout life. Thus, the growth strategy generates an ontogenetic profile of brain and body size. We consider that the growth strategy evolves by natural selection, and study its evolution using standard evolutionary-invasion analysis; that is, we consider the increase in frequency by selection of rare genetic mutations that control the growth strategy. There is a stable monomorphic female brain size in the population when rare mutants of the growth strategy cannot invade the population; that is, such resident growth strategy is 'uninvadable'[31,32]. We obtain an uninvadable growth strategy using evolutionary-invasion analysis for function-valued traits, since the growth strategy is a function of time (age). Because skill level depends (though not exclusively) on brain size due to energy conservation principles[24], the evolution of brain size causes the evolution of skill level. Accordingly, a cooperating partner's skill level and the difficulty of competitive challenges are evolving environments, which constitute the ultimate distinction between ecological and social challenges in our analysis. This evolving environment implements the notion that sociality can yield evolutionary arms races in cognition as proposed by social hypotheses[8–10].

Energy-extraction efficiency. An important quantity in the model is the individual's EEE, defined as the rate of energy extraction divided by the rate of energy extraction if the individual is maximally successful at energy extraction. We model the individual's EEE as a function of her skill level and that of cooperating or competing peers. To do this, we consider two mathematical functions commonly used in contest models: a 'power competence' function that allows for strongly decelerating EEE as the individual gains skills when she is young, and an 'exponential competence' function that allows for weaker deceleration (Fig. 2b and Extended Data Fig. 1c). We also let the skills of cooperating partners interact in an additive, multiplicative or submultiplicative (geometric mean) way (the geometric mean is a good descriptor of the average skill in the pair if peers have disparate skill levels). Additionally, we assume that if a sufficiently young individual fails to overcome a challenge, then she can extract energy from an environment facilitated by her mother.

Parameters. The model has 4 basic parameters, collectively denoted by $\boldsymbol{P}$, that specify the proportion of each social challenge, and the effects of which we study here; 13 further parameters, collectively denoted by $\boldsymbol{Q}$, that measure the metabolic costs of the brain and other tissues, the size of the brain and other tissues at birth and the demography of the population, for which empirical estimates are available; and a final 9 parameters, collectively denoted by $\boldsymbol{R}$, that measure skill metabolic costs, maternal provisioning, mutation size and how skill level affects energy extraction, for which we use reasonable values given the available data (Fig. 2c; Supplementary Information 4). For example, $\boldsymbol{R}$ parameters include the metabolic cost of memory and the values we use for this (in megajoules per year per terabyte) fall within an empirically estimated range for resting energy consumption for stored motor patterns in cerebellum Purkinje cells in rats[33]. The exact values used for $\boldsymbol{R}$ are chosen within such reasonable ranges as they yield a high ontogenetic fit between predicted and observed body and brain mass in *H. sapiens* when there are only ecological challenges (that is, $P_1 = 1$; Extended Data Fig. 3g, h). This approach is a reasonable starting point given that the fundamental constraint for a large brain is thought to be the metabolic costs of brain, which are incorporated in the estimated $\boldsymbol{Q}$ parameters. The values chosen for the $\boldsymbol{R}$ parameters mean that the difficulty of ecological challenges is high but not exceedingly so, memory is metabolically expensive (although in the low end of the empirically estimated range), and skills are moderately effective at overcoming the challenges. Using these $\boldsymbol{Q}$ and $\boldsymbol{R}$ parameter values, it was previously shown that ecological challenges alone can generate adult brain and body sizes of ancient human scale: of late *H. erectus* scale with strongly decelerating EEE and of Neanderthal scale with weakly decelerating EEE[24]. Here we use the same $\boldsymbol{Q}$ and $\boldsymbol{R}$ parameter values to study the effects of the social-challenge parameters $\boldsymbol{P}$.

Key equations. We assume that the population is large and mostly constituted by individuals with a resident growth strategy and by vanishingly rare individuals with a mutant growth strategy. At age t, a focal mutant individual has a mass of tissue i (for $i \in \{\text{b, r, s}\}$ for brain, reproductive and somatic) of $x_i(t)$ (in kilograms) and a skill level of $x_k(t)$ (in terabytes). The growth rate of tissue $i \in \{\text{b, r, s}\}$ is

$$\dot{x}_i(t) = \frac{1}{E_i} u_i(t) \left[B_{\text{rest}}(t) - \sum_{i \in \{\text{b,r,s}\}} B_i x_i(t) \right] \tag{1}$$

where $\dot{x}_i(t)$ denotes the derivative of $x_i(t)$ with respect to t. The term in square brackets is the growth-metabolic rate ($B_{\text{syn}}(t)$), which equals the resting metabolic rate, $B_{\text{rest}}(t)$, minus the maintenance metabolic rate, $\sum_{i \in \{\text{b,r,s}\}} B_i x_i(t)$. The metabolic cost of producing (respectively, maintaining) a mass unit of tissue i is E_i (respectively, B_i). The growth strategy is the fraction of the growth-metabolic rate due to the production of each tissue throughout life, and for the mutant it is denoted by $u_i(t)$ for all t and all $i \in \{\text{b, r, s}\}$ (or u for short). We let the growth strategy be the evolving trait. In turn, the mutant skill growth rate is

$$\dot{x}_k(t) = \frac{1}{E_k} [s_k B_{\text{rest,b}}(t) - B_k x_k(t)] \tag{2}$$

The brain metabolic rate is $B_{\text{rest,b}}(t)$, and the metabolic cost of gaining (respectively, maintaining) a skill unit is E_k (respectively, B_k). The fraction of brain metabolic rate allocated to energy-extraction skills is s_k. Resting metabolic rate is a power function of body mass, $x_B(t)$, and a function of EEE, which we denote by $e(x_k(t), y_k(t))$:

$$B_{\text{rest}}(t) = K e(x_k(t), y_k(t)) x_B^{\beta}(t) \tag{3}$$

where $y_k(t)$ is the skill level at age t of a resident individual. The brain metabolic rate is the sum of brain's maintenance and growth metabolic rates:

$$B_{\text{rest,b}}(t) = B_b x_b(t) + E_b \dot{x}_b(t) \tag{4}$$

An uninvadable growth strategy $u_i^*(t)$ for all t and all $i \in \{\text{b, r, s}\}$ (or u^*) is a best response to itself (similar to a Nash equilibrium) regarding the lifetime number of offspring it yields[31,32] (see 'Evolutionary differential game'). We denote the tissue mass and skill level resulting from an uninvadable growth strategy as $x_i^*(t)$ for all t and all $i \in \{\text{b, r, s, k}\}$ (or x^* for short).

Switching times. With the parameter values that we use, the uninvadable growth strategies typically produce a life history with four critical ages at which the growth strategy changes suddenly (called switching times in optimal control terminology): the age of brain growth onset t_{b0}, which is when allocation to brain growth starts; the age of brain growth arrest t_b, when allocation to brain growth stops; the age at maturity t_m, when allocation to growth of reproductive tissue starts; and the age at adulthood t_a, when allocation to growth of non-reproductive tissues stops. These four ages are an output, not parameters, of the model.

Asymptotic skill level. In adulthood (that is, after t_a), brain growth is absent and when memory is expensive enough skill growth asymptotically ceases[24]. Specifically, $\dot{x}_b^*(t) = 0$ for $t \geq t_a$ and $\dot{x}_k^*(t)$ tends to 0 as t tends to T, where T is the age of menopause. Substituting this and equation (4) into equation (2) yields the asymptotic skill level

$$\hat{x}_k = s_k \frac{B_b}{B_k} x_b^*(t_a) \tag{5}$$

Equations for EEE. In Supplementary Information 2.1, we show that EEE can be written as

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

$$e(x_k(t), y_k(t)) = \sum_{j=1}^{4} P_j \left[\frac{c_j}{c_j + d_j} + \frac{d_j}{c_j + d_j} \varphi \right] \quad (6)$$

The term in square brackets (denoted by e_j) gives the EEE when facing a challenge of type j and is composed of two terms. The first term is the proportion of time that the individual succeeds at the challenge, and the second term is the proportion of time that the individual fails at the challenge but extracts energy from maternal provisioning. The individual's competence at a type-j challenge is c_j and in general depends on her own and on her social partners' skill level. The difficulty of a type-j challenge is d_j and in general depends on the social and non-social environment (that is, on the social partners' skill level and the constant environment). Because the mutant is rare, her social partners are residents. The EEE from maternal provisioning is φ, which decreases with age.

Using our assumption of domain-general skills, we let the competence function be independent of the challenge type, $c_j(G_j(x_k, y_k)) = c(G_j(x_k, y_k))$, where $G_j(x_k, y_k)$ is a production function describing how the skills of the cooperating partners interact (that is, in an additive, multiplicative or submultiplicative way). We consider two forms for the competence function:

$$c(G_j(x_k, y_k)) = \begin{cases} G_j^{\gamma}(x_k, y_k) & \text{with power competence} \\ \exp(G_j(x_k, y_k))^{\gamma} & \text{with exponential competence} \end{cases} \quad (7)$$

where γ measures the effectiveness (decidability) of skills at the challenge. The production function $G_j(x_k, y_k)$ is

$$G_j(x_k, y_k) = \begin{cases} x_k & \text{for } j \in \{1, 3\} \\ x_k + y_k & \text{for } j \in \{2, 4\} \text{ with additive cooperation} \\ x_k \times y_k & \text{for } j \in \{2, 4\} \text{ with multiplicative cooperation} \\ \sqrt{x_k \times y_k} & \text{for } j \in \{2, 4\} \text{ with sub–multiplicative cooperation} \end{cases} \quad (8)$$

The difficulty d_j of a challenge depends on the challenge type. For an ecological or a cooperative challenge, the challenge difficulty is α, which depends on the ecological environment, which we assume is constant (this assumption can be relaxed in future extensions; Supplementary Information 9). In turn, the difficulty of a competitive challenge depends on the skill level of the individual's competitors. Since the mutant is rare, a mutant's competitors are residents, so the difficulty of a competitive challenge is the competence of the resident, $c(G_j(y_k, y_k))$. In general, the difficulty of a type-j challenge is

$$d_j(y_k) = \begin{cases} \alpha & \text{for } j \in \{1, 2\} \\ c(G_j(y_k, y_k)) & \text{for } j \in \{3, 4\} \end{cases} \quad (9)$$

We let the EEE from maternal provisioning when the individual is of age t be

$$\varphi(t) = \varphi_0 \exp(-\varphi_r t) \quad (10)$$

where φ_0 is the EEE from maternal provisioning at birth and φ_r measures the rate of decrease of maternal provisioning. The resulting equations for e_j for all cases considered can be found in Supplementary Information 2.3.

Evolutionary differential game. Let $R_0(u, v)$ be the expected lifetime number of offspring of a mutant with growth strategy u when the resident growth strategy is v. We assume that the population is kept at a constant size due to density-dependent competition through fertility rather than survival. Hence, an uninvadable growth strategy u^* maximizes the mutant's expected lifetime number of offspring when u^* is resident[34]. That is, an uninvadable growth strategy u^* satisfies

$$u^* \in \underset{u \in U}{\operatorname{argmax}} R_0(u, u^*) \quad (11)$$

where U is the set of feasible growth strategies. Assuming that the mortality rate μ is constant (which can be relaxed in future extensions; Supplementary Information 9) and that reproductive tissue is narrowly defined so that it is not involved in offspring maintenance (for example, defined as preovulatory ovarian follicles), the mutant's expected lifetime number of offspring when v is resident is

$$R_0(u, v) \propto \int_0^T \exp(-\mu t) x_r(t) \mathrm{d}t \quad (12)$$

Thus equation (11) poses a differential game problem: it is a 'game' between mutant and resident because the mutant's payoff ($R_0(u, v)$) depends on the resident strategy, it is 'differential' because it depends on differential equations (equations (1) and (2)), and it is 'evolutionary' rather than a typical differential game because only the mutant's payoff is maximized rather than both the mutant and resident's payoffs. Because equation (11) involves maximization with respect to functions (u) rather than points, this maximization poses an optimal control problem. We solve this problem numerically by finding a best response to the resident (optimal control problem), setting the best response as the resident, and iterating until convergence to a point at which the mutant and resident strategies are indistinguishable to a chosen extent. To do so, we use the software GPOPS[35].

Figure specifications. For Fig. 3a, b, plots are around only ecological challenges; that is, for a given plot, the remaining two P_j's are set to zero. For social challenges, the arrows in Fig. 3c describe the qualitative effect determined in Fig. 3a, b of increasing the proportion of a social challenge as the proportion of ecological challenge decreases; for ecological challenges, the arrows describe the qualitative effect of increasing the environmental difficulty α as found in Extended Data Fig. 3g, h. The patterns in Fig. 3a–c also hold around the best-fitting $\boldsymbol{P}^*$ for *H. sapiens*; that is, when for a given plot, the remaining two P_j's are set to the values of $\boldsymbol{P}^*$ (Extended Data Fig. 4b, c). The 'missing' dots in Fig. 3d are P_j combinations that did not converge to an uninvadable growth strategy (for example, owing to cycling solutions, suggesting possible evolutionary branching (female dimorphism) in brain size) or that were unreachable from lack of convergence of nearby runs (Supplementary Information 5). For Fig. 3a–e, cooperation is submultiplicative and for Fig. 3d, e, competence is exponential (see Extended Data Fig. 4 for all cases).

Figure 4a shows the hominin species for which we find a near-perfect adult fit (that is, for which the best adult fit is greater than the chosen threshold of $-D(\tau_a) = -0.05$; Extended Data Figs. 6–8). For Fig. 4a, cooperation is submultiplicative (respectively additive) for weakly (respectively strongly) decelerating EEE. In Fig. 4b, dots are the values for an average *H. sapiens* female as previously reported[21]. The resulting life periods in Fig. 4b are defined as 'childhood', when there has not been allocation to production of reproductive tissue from birth; 'adolescence', when there is allocation to production of somatic and reproductive tissues; and 'adulthood', when there is only allocation to production of reproductive tissue. The EEE from maternal provisioning at birth (part of the $\boldsymbol{R}$ parameters) in Fig. 4b is slightly smaller than its benchmark value to improve ontogenetic fit further without affecting adult fit (ontogenetic fit is $-E(D(\tau)) = -0.22$ using $\varphi_0^* = 0.5$ rather than -0.33 using the benchmark $\varphi_0 = 0.6$; Supplementary Information 6, 8).

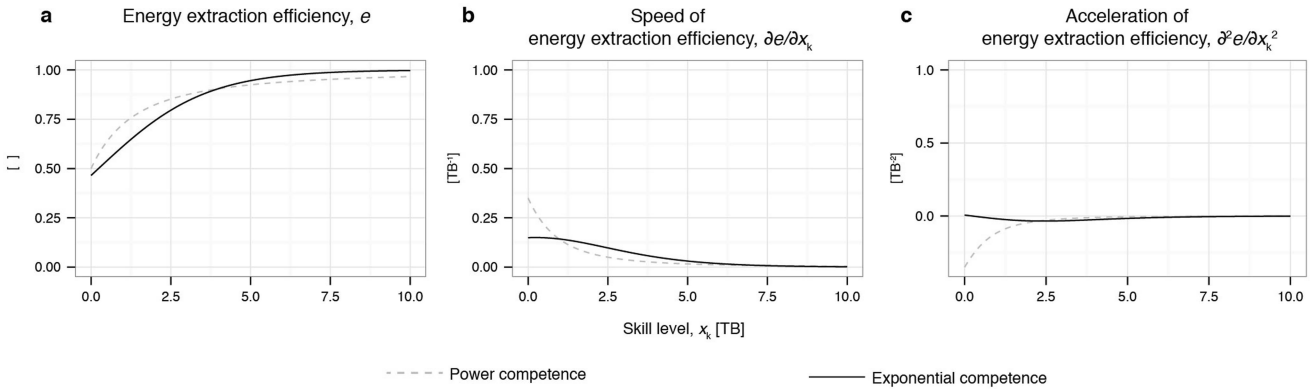
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Code that supports the findings of this study is available in Zenodo with the identifier https://doi.org/10.5281/zenodo.1197479.

Data availability. Data of predicted brain size, body size and skill level for the various challenge combinations as generated by this study and as used for Figs. 3, 4 and Extended Data Fig. 4, 6–8 have been deposited in Zenodo with the identifier https://doi.org/10.5281/zenodo.1197479. Complete numerical solutions, including growth strategies across the parameter sweep totalling 200 GB of data have been deposited in Zenodo with the identifier https://doi.org/10.5281/zenodo.1217123.

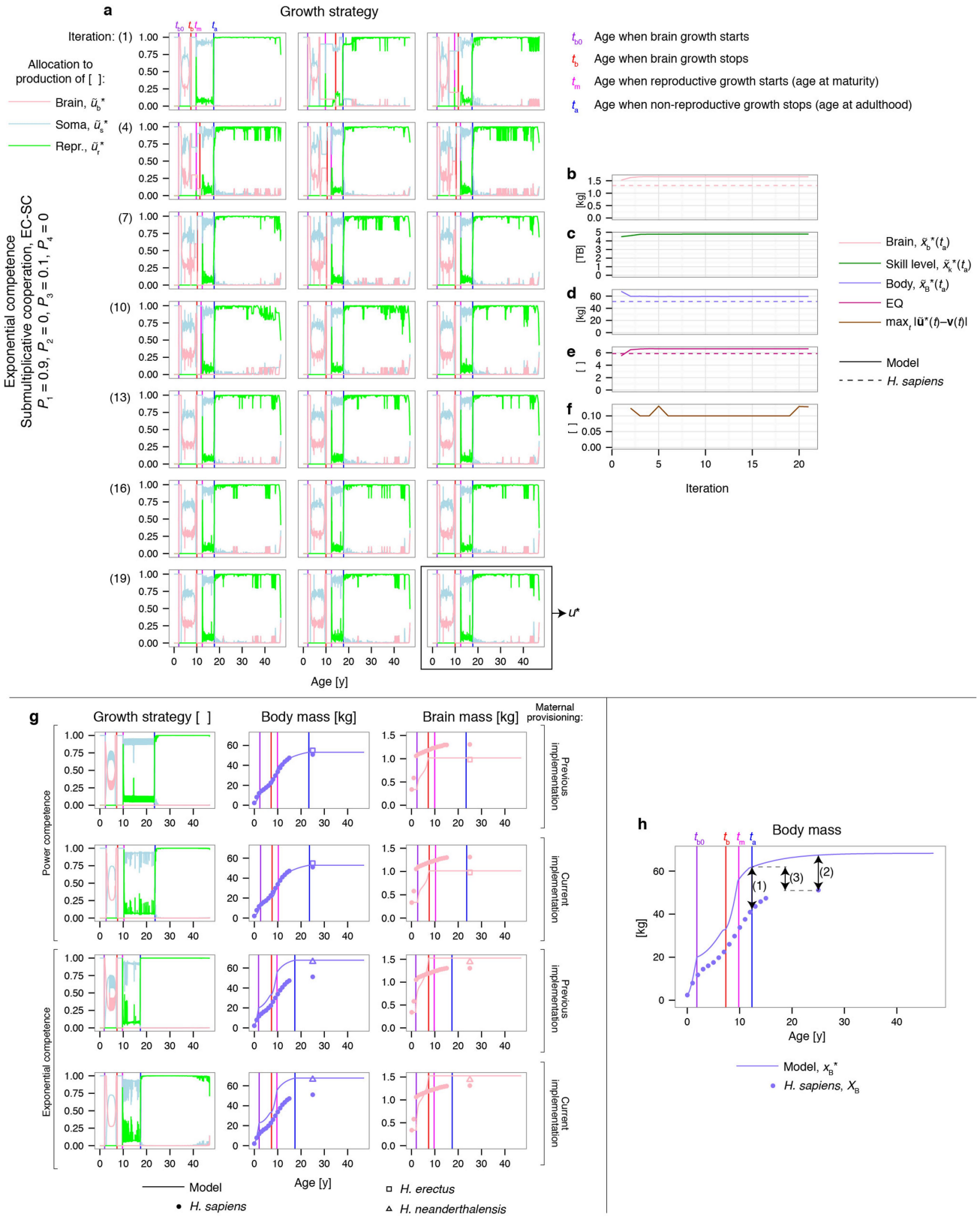
31. Lehmann, L., Mullon, C., Akçay, E. & Van Cleve, J. Invasion fitness, inclusive fitness, and reproductive numbers in heterogeneous populations. *Evolution* **70**, 1689–1702 (2016).
32. Lion, S. Theoretical approaches in evolutionary ecology: environmental feedback as a unifying perspective. *Am. Nat.* **191**, 21–44 (2018).
33. Howarth, C., Peppiatt-Wildman, C. M. & Attwell, D. The energy use associated with neural computation in the cerebellum. *J. Cereb. Blood Flow Metab.* **30**, 403–414 (2010).
34. Mylius, S. D. & Diekmann, O. On evolutionarily stable life histories, optimization and the need to be specific about density dependence. *Oikos* **74**, 218–224 (1995).
35. Patterson, M. A. & Rao, A. V. GPOPS-II: a MATLAB software for solving multiple-phase optimal control problems using hp-adaptive Gaussian quadrature collocation methods and sparse nonlinear programming. *ACM Trans. Math. Softw.* **41**, 1–37 (2014).
36. Schniter, E., Gurven, M., Kaplan, H. S., Wilcox, N. T. & Hooper, P. L. Skill ontogeny among Tsimane forager-horticulturalists. *Am. J. Phys. Anthropol.* **158**, 3–18 (2015).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 1 | Shape of EEE versus skill. a–c, Plots of EEE (e), its speed and acceleration with respect to skill level under power competence and exponential competence with only ecological challenges (that is, $P_1 = 1$) for the parameter values used and without maternal provisioning (that is, $\varphi = 0$, so that $e = S_1$). For comparison, the curves for power competence are displaced to the left by 1 unit because skill level at birth is 1 for power competence but 0 for exponential competence. **a**, **b**, e and its speed at birth and during young ages are smaller for exponential competence than for power competence. **c**, However, the acceleration in e at birth and at young ages is larger for exponential competence than for power competence.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

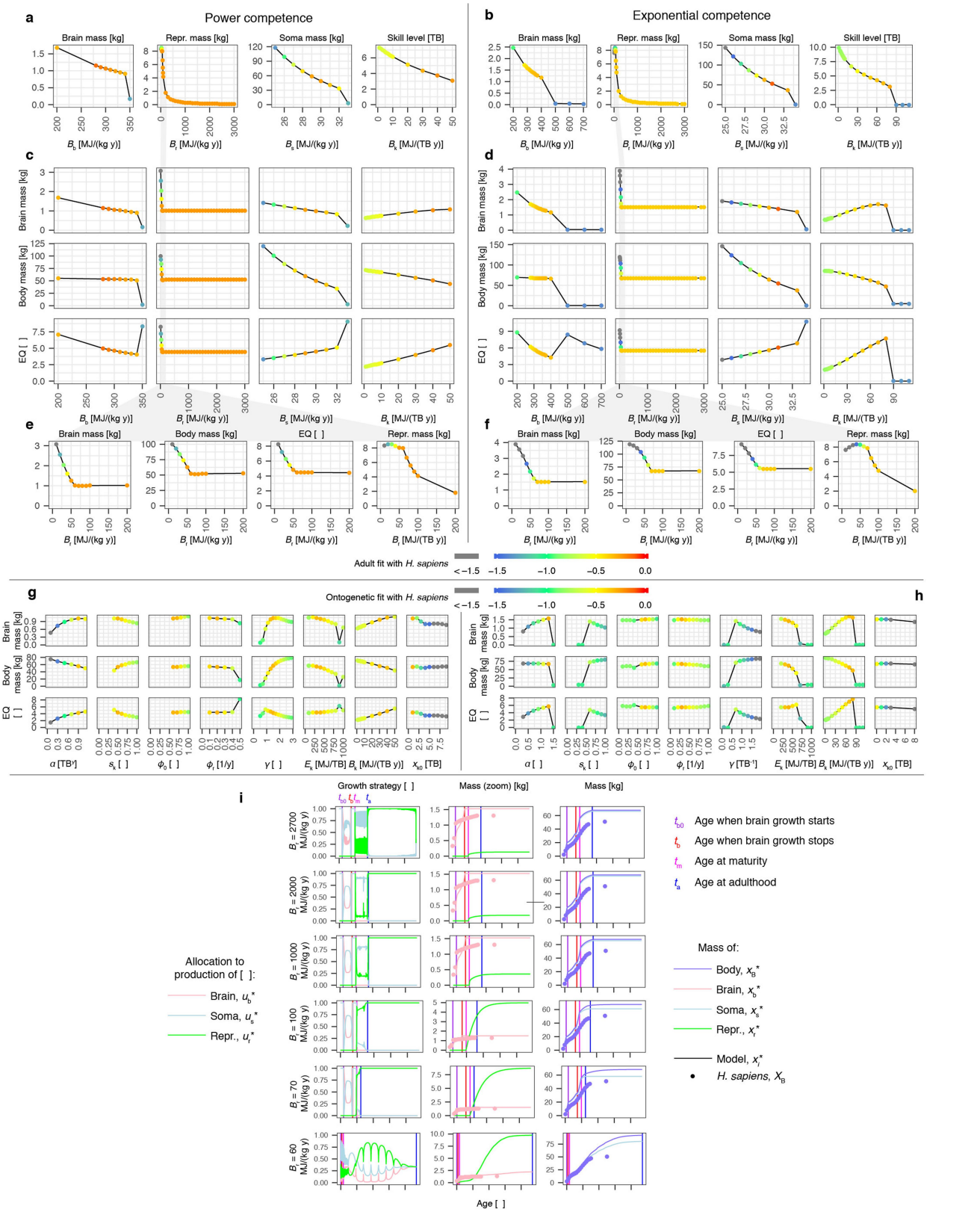


Extended Data Fig. 2 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 2 | Method implementation. a, Typical result with convergence to an uninvadable growth strategy. For the *i*th best-response iteration, the growth strategy shown is the resident (v) whose best response ($\tilde{u}^*$) is shown next, which is the resident of the $i+1$th iteration. Convergence to a best response to itself (u^*) was declared visually, in this case, at iteration 21. **b–f**, Reporting variables across the best-response iterations in **a**. **b–e**, Resulting adult body mass, brain mass, skill level and encephalization quotient across iterations. These values tend to converge more quickly than the growth strategy (**a**). **f**, Rather than visually declaring convergence, convergence should ideally be declared when the difference between mutant and resident is below a chosen threshold. However, numerical jittering prevented the use of this criterion. For example, **f** shows the maximum of $|\tilde{u}^*(t)-v(t)|$ across t for each iteration in **a**. Without numerical jittering, this maximum should decrease as the growth strategy approaches a best response to itself. However, numerical jittering causes this maximum to be at least equal to the maximum mutation size $\delta=0.1$. The maximum of $|\tilde{u}^*(t)-v(t)|$ is occasionally greater than δ because $\tilde{u}^*$ and v have different partitions over t and we use the following approximation: for each t in the t partitioning of $\tilde{u}^*$, we find the closest t in the t partitioning of $\boldsymbol{v}$ and calculate the difference $|\tilde{u}^*(t)-v(t)|$ at these relatively close times; this may occasionally cause the difference to be the larger than δ when strategies change suddenly with t. Alternative measures of convergence were similarly inadequate (for example, $\Sigma_t\ |\tilde{u}^*(t)-v(t)|$). **g**, We implement maternal provisioning differently than before[24] to incorporate it when there are social challenges. The difference yields no detectable difference in predicted brain and body mass with only ecological challenges after slightly adjusting the EEE from maternal provisioning of a newborn (φ_0): before[24], $\varphi_0=0.6$ for power and $\varphi_0=0.8$ for exponential competence were used; here, $\varphi_0=0.4$ for power and $\varphi_0=0.6$ for exponential competence were used. **h**, Three ways to measure adult fit: (1) at the predicted age of adulthood $(x^*_B(t_a)-X_B(t_a))$; (2) at the observed age of adulthood $(x^*_B(\tau_a)-X_B(\tau_a))$; and (3) at the predicted age of adulthood for the prediction and at the observed age of adulthood for the observation $(x^*_B(t_a)-X_B(\tau_a))$. We use option 2.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

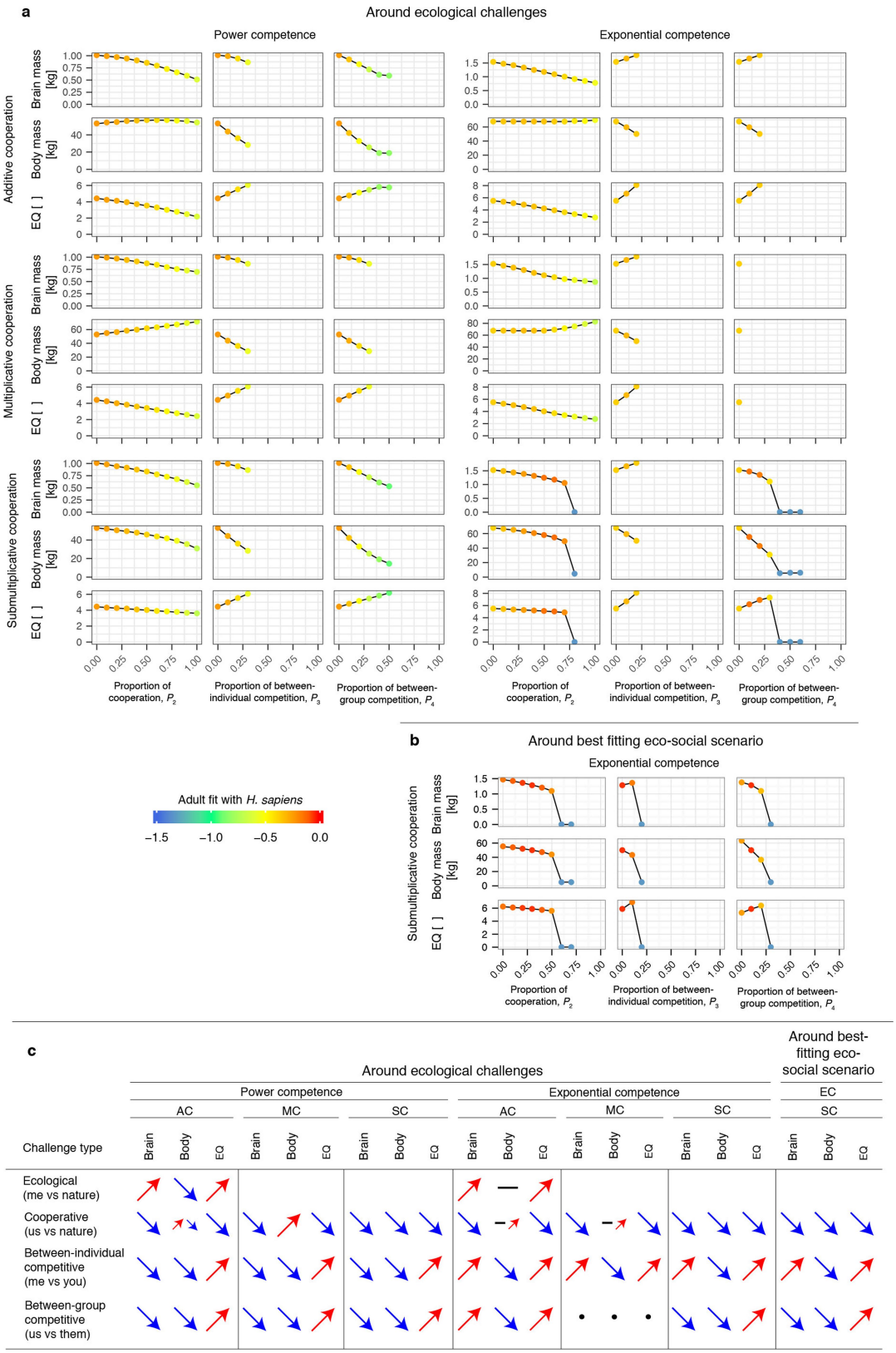


Extended Data Fig. 3 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

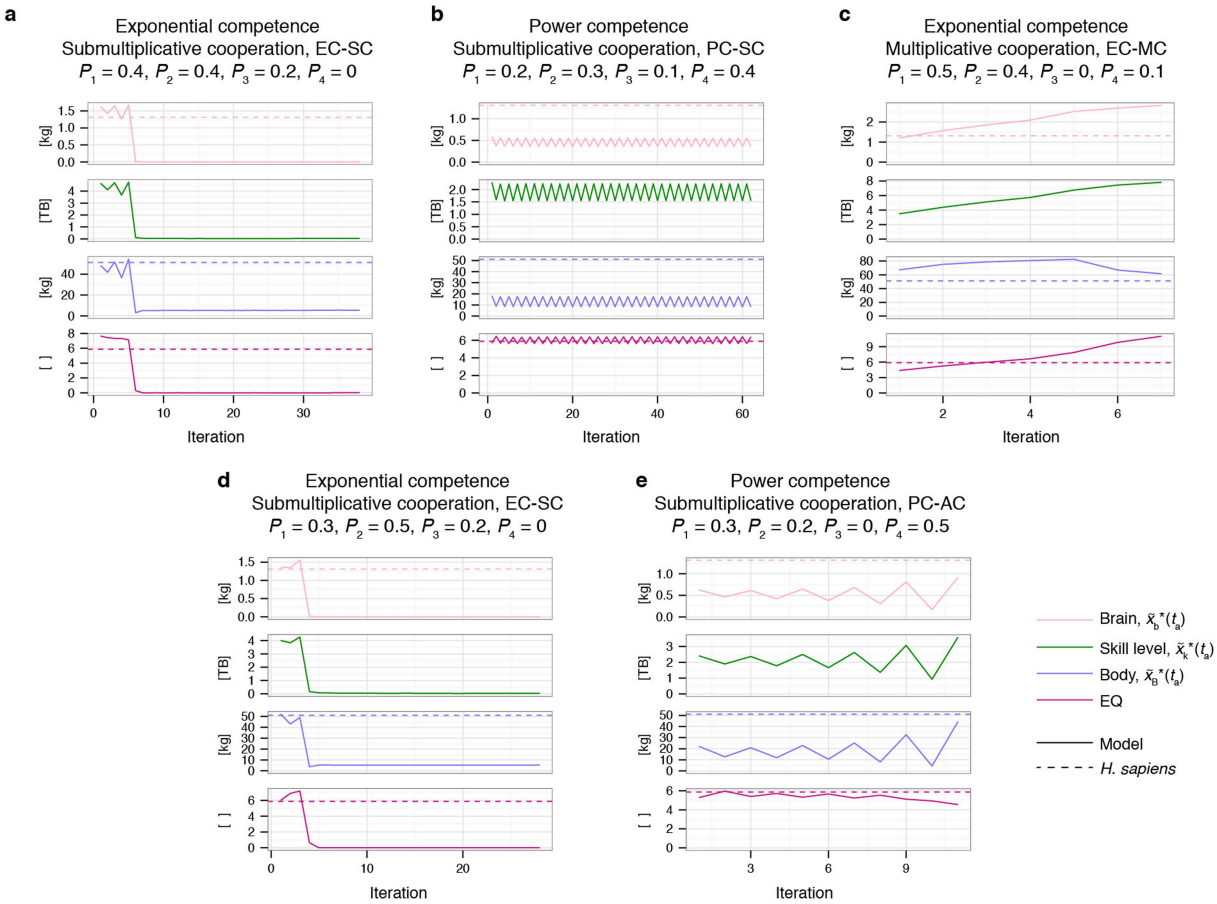
Extended Data Fig. 3 | Effects of *Q* and *R* parameters. a, **b**, Effects of maintenance costs (B_i) on the corresponding tissue mass or skill level. Each B_i tends to decrease the value $x_i^*(\tau_a)$ for the corresponding i, but not necessarily for the other i (see **c**, **d**). **c**, **d**, Effect of B_i on adult brain mass, body mass and encephalization quotient. With power competence (**c**), when $B_b = 310$ and 340 MJ kg^{-1} per year (y), the predicted adult brain mass is $x_b^*(\tau_a) = 1.0298$ and 0.9133 kg, respectively. With exponential competence (**d**), when $B_b = 310$, 340 and 370 MJ kg^{-1} y^{-1}, the predicted adult brain mass is $x_b^*(\tau_a) = 1.542$, 1.3973 and 1.2767 kg, respectively. **e**, **f**, Effects of B_r when B_r is small. When B_r varies between 70 and 2,700 MJ kg^{-1} y^{-1}, B_r has no detectable effect on adult brain mass and encephalization quotient. **g**, **h**, Ontogenetic fit with *H. sapiens* around the used values for each of the ***R*** parameters (except δ). The ontogenetic fit is approximately maximized around the benchmark values chosen previously[24], which are also used here (except for φ_0 given our improved implementation of φ). **i**, Effect of B_r on the predicted life history with exponential competence. In the left column, from top to bottom, as B_r decreases, the allocation to the growth of reproductive tissue during adolescence increases (u_r^* between t_m and t_a) and adolescence shortens. In the central column, the increased allocation to the growth of reproductive tissue increases the mass of reproductive tissue, but brain mass does not change with B_r for $B_r \geq 70$ MJ kg^{-1} y^{-1}. In the right column, as the mass of reproductive tissue increases, body mass increases slightly, which is more noticeable for $B_r \leq 100$ MJ kg^{-1} y^{-1}. An exceedingly small B_r (<70 MJ kg^{-1} y^{-1}) disrupts the predicted life history, which with $B_r = 60$ MJ kg^{-1} y^{-1} is severely different from that of *H. sapiens* (for example, there is brain growth late in life and reproductive growth from birth). Similar results arise for even smaller B_r. In **a**–**i** there are only ecological challenges and we use the previous[24] definition of φ.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



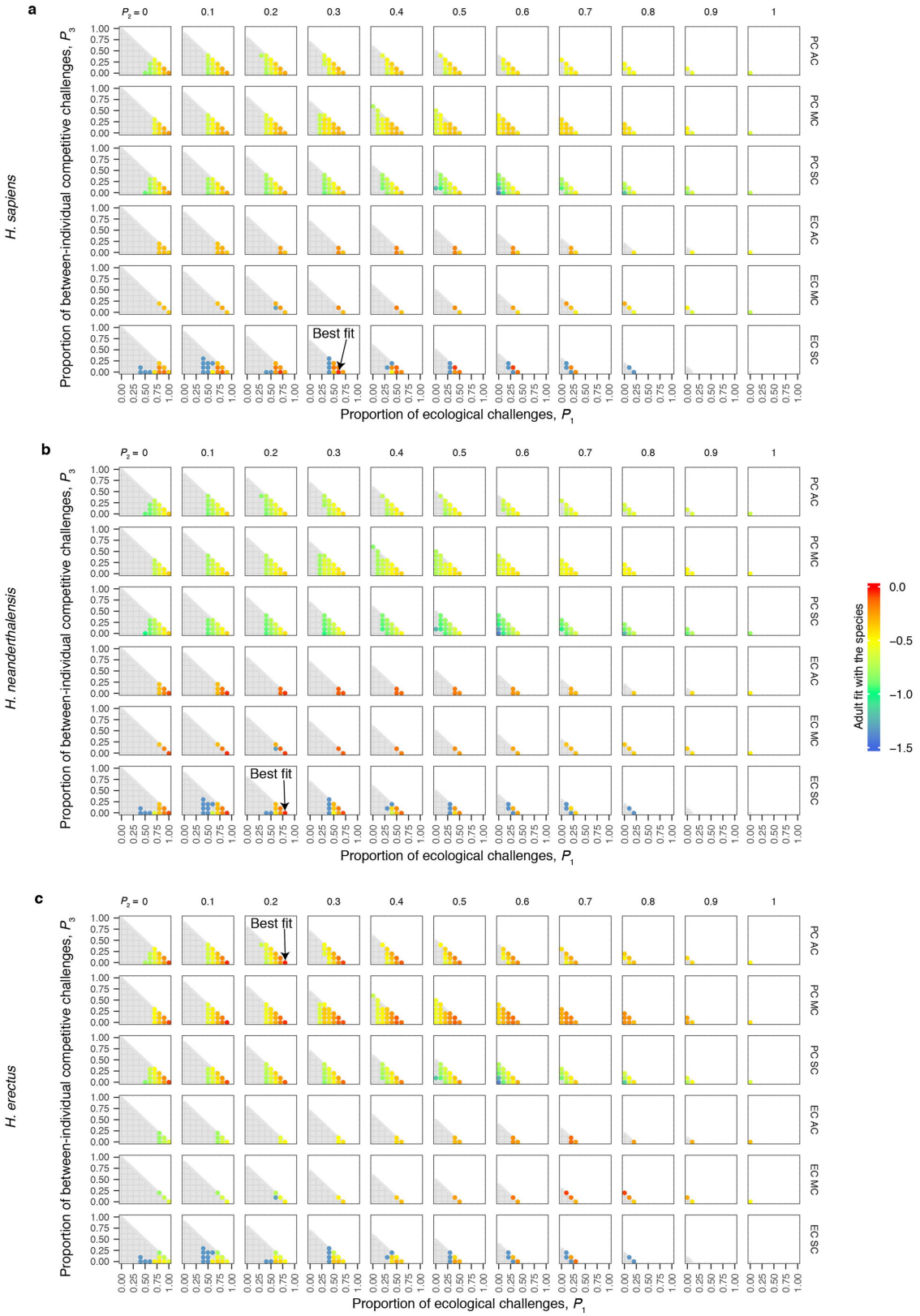
Extended Data Fig. 4 | Effects of challenge types on brain size. **a**, **b**, Outer rows are for the cooperation cases that were considered; outer columns are for the competence cases. **a**, Around the pure ecological scenario (that is, in a given plot for P_j as P_1 decreases, the remaining two P_j's are set to zero). **b**, Around the best fitting scenario for *H. sapiens* (that is, in a given plot for P_j as P_1 decreases, the remaining two P_j's are set to the best fitting $\boldsymbol{P}^*$ found in Fig. 3d. **c**, Summary of the qualitative effects of challenge types on brain size. For social challenges, the direction of the arrows is taken from **a**, **b**. For ecological challenges, the direction of the arrow is taken from Extended Data Fig. 3g as the environmental difficulty α increases. A dash (−) indicates an approximately invariant relationship and a dot (·) indicates insufficient data points for identifying a relationship. The arrows in Fig. 3c are taken from this summary, in which, for social challenges, the arrows are those of submultiplicative cooperation. AC, additive cooperation; EC: exponential competence; MC, multiplicative cooperation; SC, submultiplicative cooperation.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



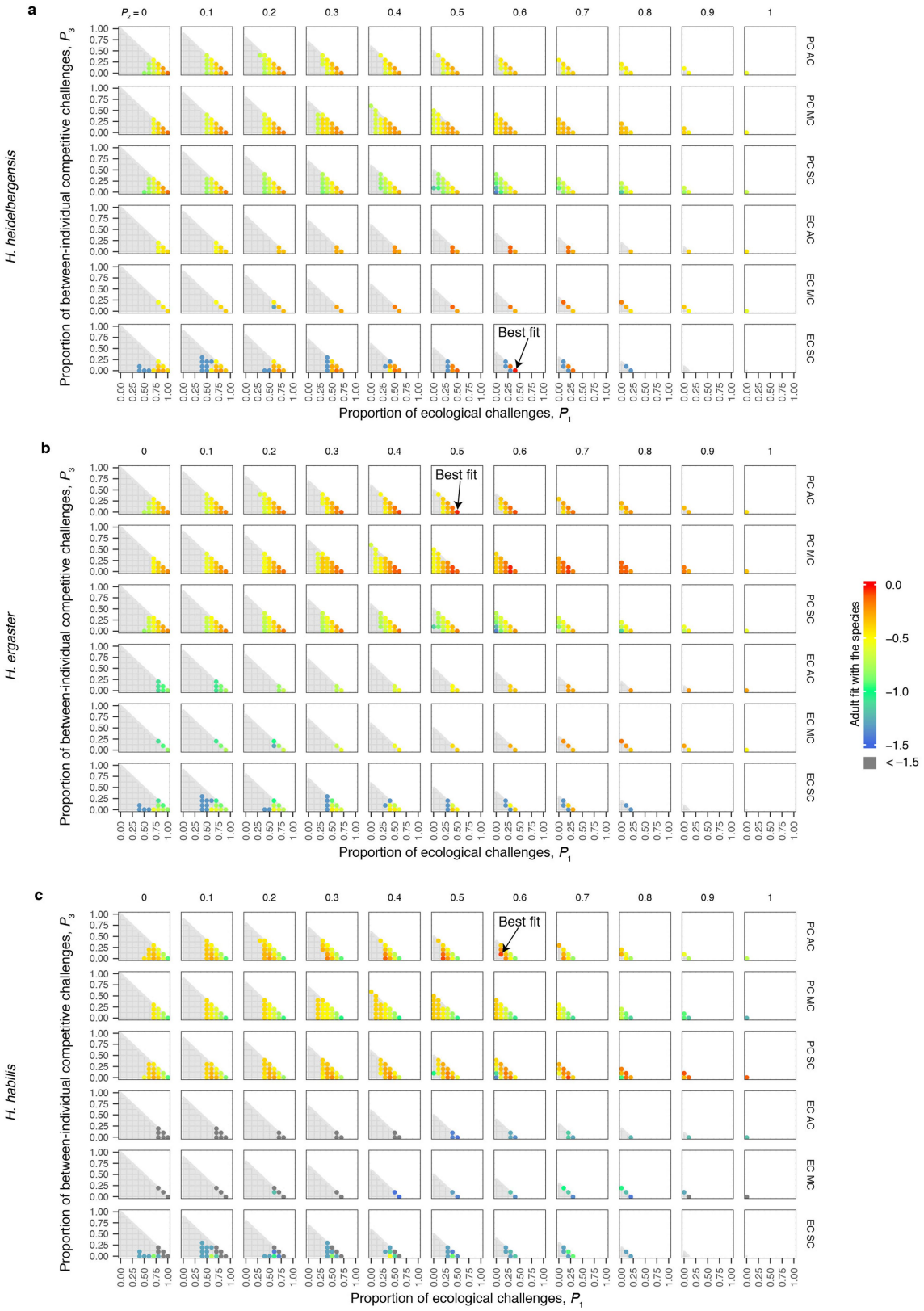
Extended Data Fig. 5 | Typical results when there is convergence to no brain growth or when there is no convergence to an uninvadable growth strategy. a–**e**, Adult values over best-response iterations for cases of no brain growth or no convergence to an uninvadable strategy. **a**, Amplifying cycle leads to no brain growth. **b**, Stable cycle. **c**, Arms race that ends when the solver warns that the optimal control problem (OCP) may be infeasible. This might arise if the best response to the last iteration necessarily involves a substantially different growth strategy, which is not allowed in the optimization as the best response is constrained to be sufficiently similar to that in the previous iteration. It is possible that such substantially different best response involves either no brain growth (for example, as seen under purely ecological challenges when the environmental difficulty is exceedingly high[24] (Supplementary Information 4.4)) or substantially more allocation to brain growth (which appears unlikely given the energetic constraints). **d**, A short arms race in encephalization quotient that leads to no brain growth. **e**, Amplifying cycle that ends when the solver warns that the OCP may be infeasible.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



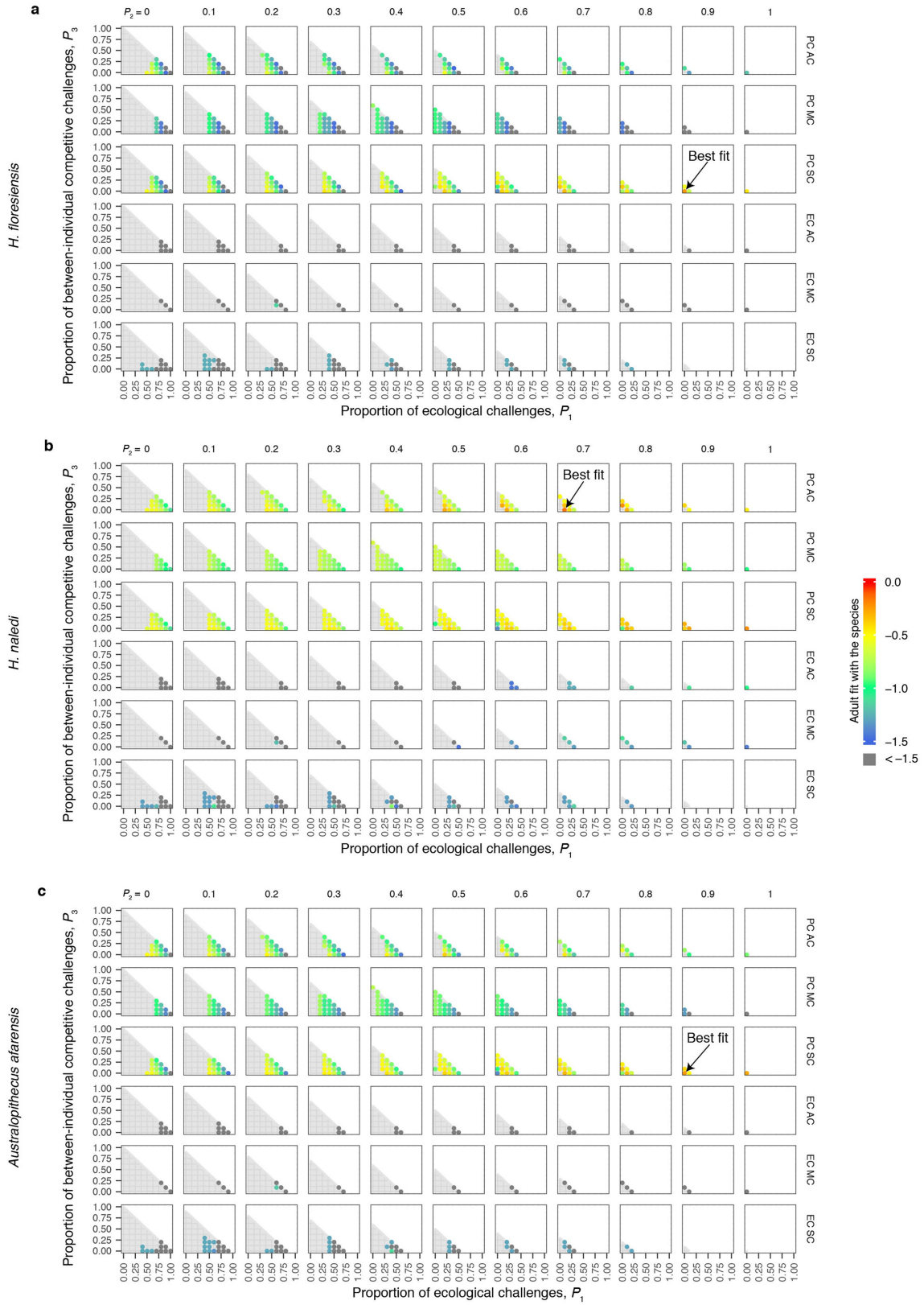
Extended Data Fig. 6 | Identification of best-fitting scenarios across hominins. Adult fit of predicted adult brain and body mass with those observed in a given species across parameter values for all cases considered. Each dot's colour gives the adult fit, $-D(\tau_a)$, for the corresponding parameter combination and case. **a**, *H. sapiens*. **b**, *H. neanderthalensis*. **c**, *H. erectus*.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



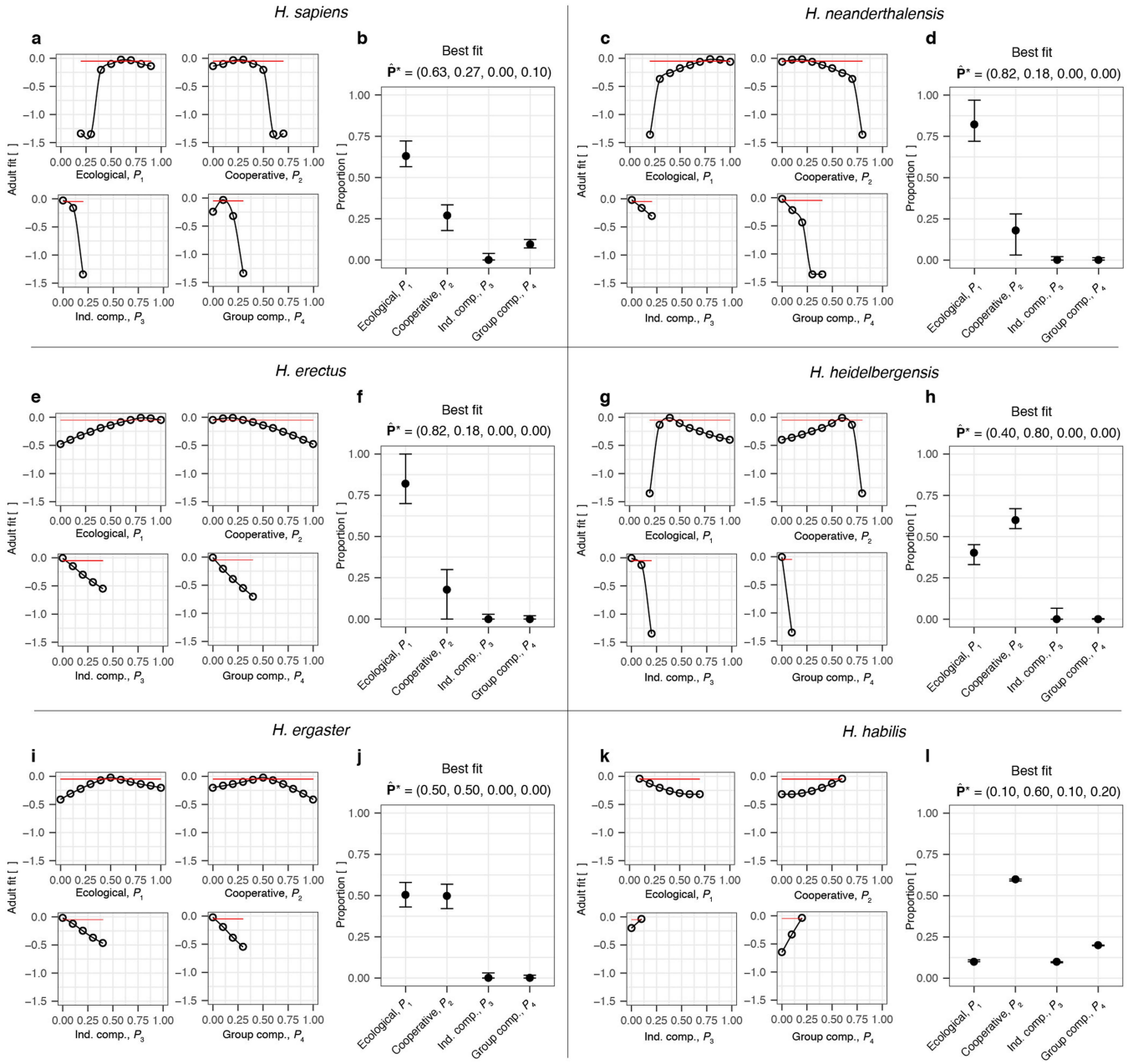
Extended Data Fig. 7 | Identification of best-fitting scenarios across hominins, continued. See legend of Extended Data Fig. 6 for details. **a**, *H. heidelbergensis*. **b**, *H. ergaster*. **c**, *H. habilis*.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



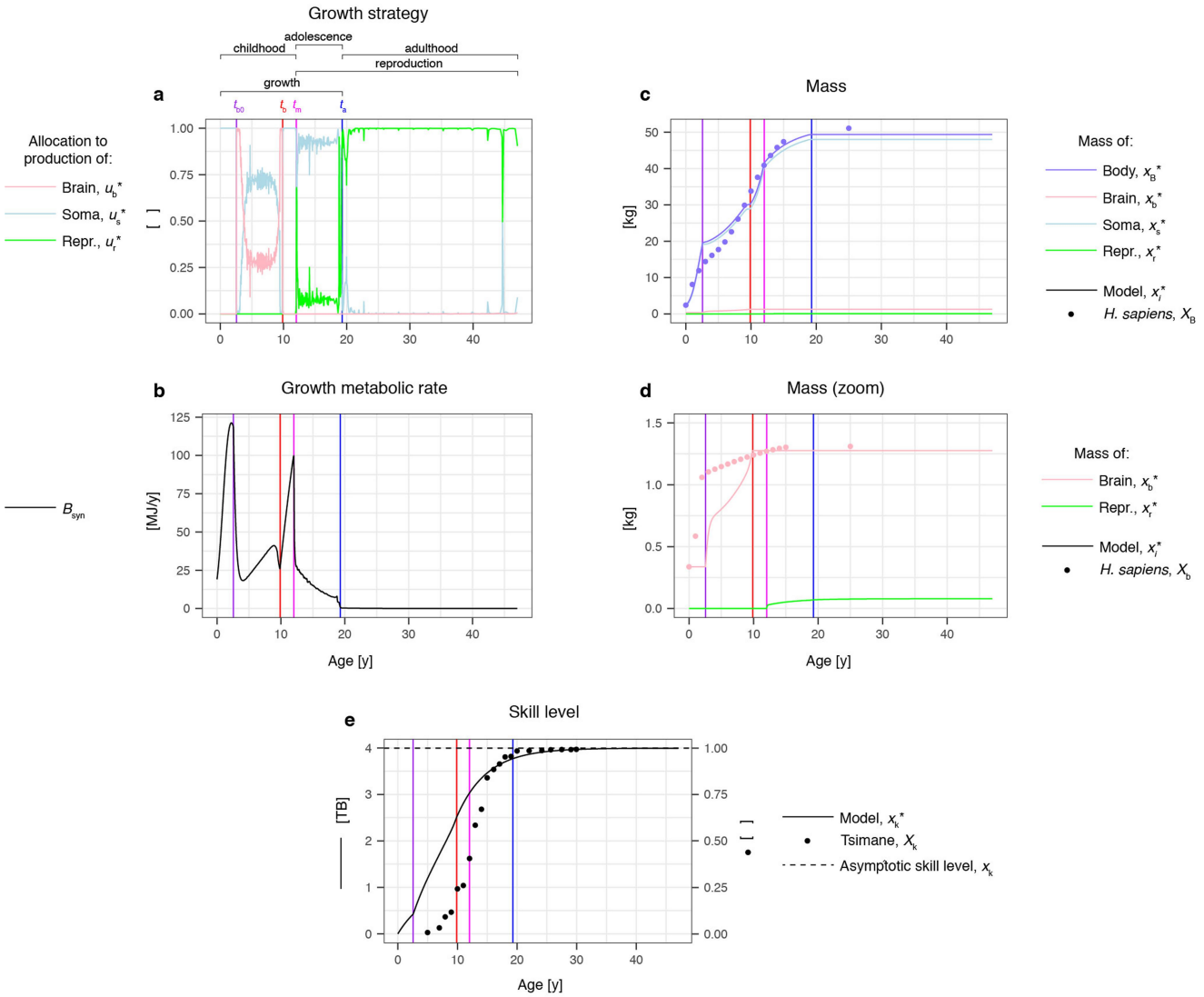
Extended Data Fig. 8 | Identification of best-fitting scenarios across hominins, continued. See legend of Extended Data Fig. 6 for details. **a**, *H. floresiensis*. **b**, *H. naledi*. **c**, *Australopithecus afarensis*. The best adult fit is −0.24 (**a**), −0.14 (**b**) and −0.23 (**c**).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | High fit intervals for best-fitting scenarios across hominins. Here we show high fit intervals around the best fitting scenarios across hominins having a best adult fit greater than -0.05. **a**, **c**, **e**, **g**, **i**, **k**, For the top left plot, as P_1 increases, P_2 decreases, whereas for the remaining plots as P_2, P_3 and P_4 increase, P_1 decreases; for a given plot, the remaining P_j are set to the corresponding $\boldsymbol{P}^*$ shown in Fig. 4a (that is, plots are around $\boldsymbol{P}^*$). The dots are the adult fit and the lines are interpolated values using a monotone Hermite spline (splinefun with method monoH.FC in R). The red line is $-D(\tau_a) = -0.05$. **b**, **d**, **f**, **h**, **j**, **l**, The whiskers are the high fit intervals where adult fit is greater than -0.05 and the dots are the estimated $\hat{\mathbf{P}}^*$ giving the best adult fit for the species in the interpolation. The cases of competence and cooperation are as found in Extended Data Figs. 6, 7. Note that for *H. habilis*, the high fit intervals may be wider as the adult fit is increasing at the end of the values of P_2, P_3 and P_4 for which uninvadable growth strategies were obtained.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Detailed life history resulting from the best-fitting scenario for *H. sapiens*. Plots correspond to Fig. 4b. **a**, The growth strategy generating the life history. **b**, The resulting growth metabolic rate. **c**, **d**, The mass of all tissues. **e**, The skill level. For comparison with the model's predicted skill level, x_k^*, the black dots in **e** are the observed cumulative distribution of self-reported acquisition ages of food production skills in female Tsimane horticulturalists[36].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0104-4

Cellular milieu imparts distinct pathological α-synuclein strains in α-synucleinopathies

Chao Peng, Ronald J. Gathagan, Dustin J. Covell, Coraima Medellin, Anna Stieber, John L. Robinson, Bin Zhang, Rose M. Pitkin, Modupe F. Olufemi, Kelvin C. Luk, John Q. Trojanowski & Virginia M.-Y. Lee*

In Lewy body diseases—including Parkinson's disease, without or with dementia, dementia with Lewy bodies, and Alzheimer's disease with Lewy body co-pathology[1]—α-synuclein (α-Syn) aggregates in neurons as Lewy bodies and Lewy neurites[2]. By contrast, in multiple system atrophy α-Syn accumulates mainly in oligodendrocytes as glial cytoplasmic inclusions (GCIs)[3]. Here we report that pathological α-Syn in GCIs and Lewy bodies (GCI-α-Syn and LB-α-Syn, respectively) is conformationally and biologically distinct. GCI-α-Syn forms structures that are more compact and it is about 1,000-fold more potent than LB-α-Syn in seeding α-Syn aggregation, consistent with the highly aggressive nature of multiple system atrophy. GCI-α-Syn and LB-α-Syn show no cell-type preference in seeding α-Syn pathology, which raises the question of why they demonstrate different cell-type distributions in Lewy body disease versus multiple system atrophy. We found that oligodendrocytes but not neurons transform misfolded α-Syn into a GCI-like strain, highlighting the fact that distinct α-Syn strains are generated by different intracellular milieus. Moreover, GCI-α-Syn maintains its high seeding activity when propagated in neurons. Thus, α-Syn strains are determined by both misfolded seeds and intracellular environments.

The diverse nature of α-synucleinopathies suggests that they may be caused by distinct α-Syn strains[4–8]. To investigate whether GCI-α-Syn and LB-α-Syn represent two distinct strains, sarkosyl-insoluble α-Syn was isolated from the brains of patients with multiple system atrophy (MSA), which exists as two subtypes—the Parkinsonian subtype (MSA-P) and the cerebellar subtype (MSA-C)[9,10]—and brains from patients with Lewy body disease (Extended Data Fig. 1a and Supplementary Tables 1, 2). First we evaluated the extent of Ser129 phosphorylation (pS129), which is a hallmark of pathological α-Syn[11,12], on GCI-α-Syn and LB-α-Syn, and found much less pS129 on GCI-α-Syn than on LB-α-Syn (Fig. 1a, b). Second, we analysed conformational differences between GCI-α-Syn and LB-α-Syn using proteinase K digestion. Proteinase K digestion shows predominantly undigested α-Syn for GCI-α-Syn (1st band in Fig. 1c), whereas LB-α-Syn was cleaved into smaller fragments (2nd–4th bands in Fig. 1c). The relative resistance of GCI-α-Syn to proteinase K digestion was further confirmed using increasing concentrations of proteinase K (Fig. 1d), which indicates that GCI-α-Syn might form a more-compact structure than LB-α-Syn. Epitope mapping showed that the 2nd band after proteinase K digestion was truncated mainly at the N terminus, whereas the 3rd and 4th bands were truncated mainly at the C terminus (Extended Data Fig. 1b, c). GCI-α-Syn and LB-α-Syn also produced distinct banding patterns when digested with trypsin or thermolysin, further demonstrating their different conformations (Extended Data Fig. 1d–g).

To confirm that GCI-α-Syn and LB-α-Syn have different conformations, we immunostained sections of diseased brains with the monoclonal antibody (MAb) Syn7015, which is selective for a synthetic α-Syn strain[13]. At low concentrations Syn7015 preferentially recognized GCIs over Lewy bodies, whereas another MAb Syn303 that detects pathological α-Syn[13,14] immunostained GCIs and Lewy bodies equally well (Extended Data Fig. 2a, b). Semi-quantitative analyses of Lewy bodies and GCIs[15,16] stained by Syn7015 or Syn303 on adjacent sections showed that Syn7015 preferentially recognized GCIs over Lewy bodies (Fig. 1e), which was also supported by the ratio of total area occupied by Syn7015-positive over Syn303-positive pathology (Fig. 1f, g and Extended Data Fig. 2c). This preferential recognition further demonstrates that there are conformational differences between GCI-α-Syn and LB-α-Syn.

To determine whether structural differences between GCI-α-Syn and LB-α-Syn influence their seeding activities, we treated primary oligodendrocytes expressing α-Syn with an equal amount of GCI-α-Syn, LB-α-Syn or α-Syn preformed fibrils (PFFs)[17]. GCI-α-Syn is much more potent than LB-α-Syn and PFFs at seeding α-Syn aggregation in oligodendrocytes (Fig. 1h, i and Extended Data Fig. 3a, b). Specifically, GCI-α-Syn is approximately 1,000-fold more potent than LB-α-Syn and PFFs (Fig. 1j): 30 ng of LB-α-Syn induced a similar amount of pathology as 30 pg of GCI-α-Syn, and 30 μg of PFFs were comparable to 30 ng of GCI-α-Syn. The purity of the oligodendrocyte cultures and the presence of α-Syn pathology in oligodendrocytes were confirmed by immunofluorescence staining (Extended Data Fig. 3c–f).

Because the high potency of GCI-α-Syn in inducing oligodendrocyte pathology is consistent with the distribution of this strain in oligodendrocytes of patients with MSA, we hypothesized that the properties of GCI-α-Syn and LB-α-Syn dictate their differential cell-type distribution in patients (hypothesis 1 in Extended Data Fig. 3g), and speculated that LB-α-Syn is more potent than GCI-α-Syn in inducing neuronal pathology. Primary neurons treated with GCI-α-Syn developed many more α-Syn inclusions than those treated with LB-α-Syn or PFFs (Fig. 2a, b and Extended Data Fig. 4a, b), and GCI-α-Syn also was about 1,000-fold more potent than LB-α-Syn or PFFs in inducing neuronal α-Syn pathology (Fig. 2c, d). Furthermore, using a QBI-293 cell line expressing human α-Syn (hereafter, 'QBI-WT-Syn' cells)[18], we confirmed that GCI-α-Syn is about 1,000-fold more potent than LB-α-Syn and PFFs (Fig. 2e, f and Extended Data Fig. 4c–e). To rule out possible contributions of contaminating proteins, GCI-α-Syn and LB-α-Syn were further purified by immunoprecipitation. Immunoprecipitation-purified GCI-α-Syn maintained its markedly higher seeding ability compared to LB-α-Syn (Extended Data Fig. 4f). Moreover, the addition of PFFs to immunoprecipitation-depleted GCI-α-Syn preparations did not increase the seeding ability of PFFs (Extended Data Fig. 4g, h).

Because GCI-α-Syn is more resistant to proteinase K digestion (Fig. 1c, d), we investigated whether the high potency of GCI-α-Syn is due to its resistance to degradation. Previously we showed that exogenously added misfolded α-Syn accumulates in lysosomes and that treating the cells with chloroquine (a lysosomal inhibitor) could increase the amount of α-Syn pathology that was induced[19]. To test this hypothesis, primary neurons seeded with GCI-α-Syn, LB-α-Syn or PFFs were treated with chloroquine. Chloroquine treatment similarly increased the pathology induced by GCI-α-Syn, LB-α-Syn and PFFs (Extended

The Department of Pathology and Laboratory Medicine, Institute on Aging and Center for Neurodegenerative Disease Research, the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. *e-mail: vmylee@upenn.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

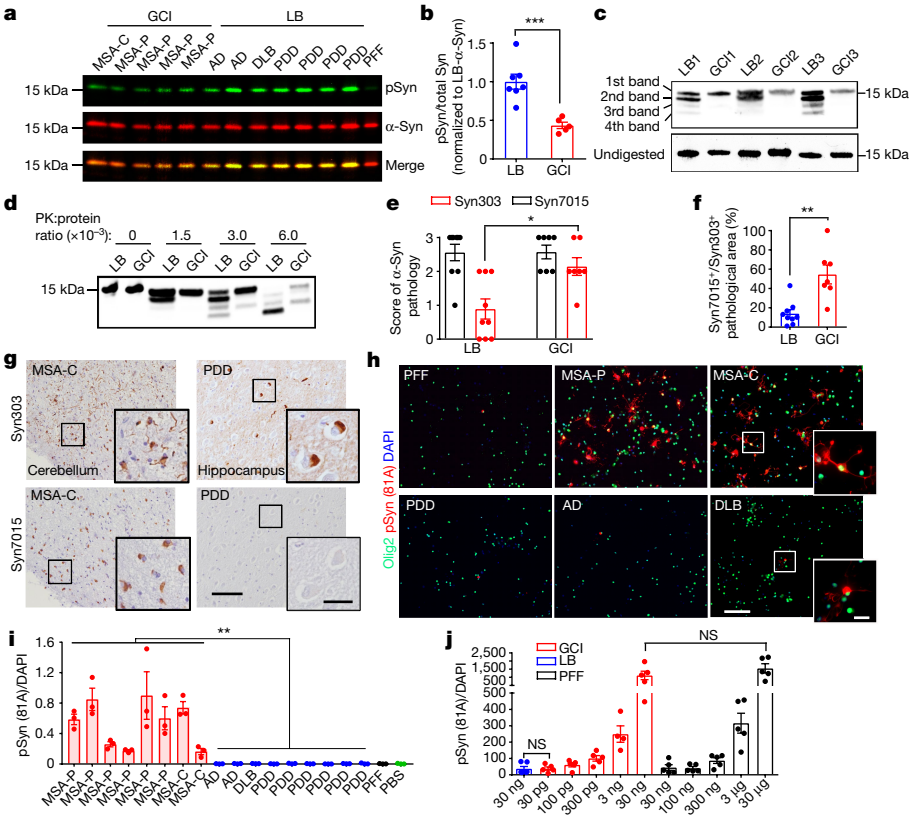


Fig. 1 | GCI-α-Syn and LB-α-Syn represent two distinct strains. **a**, GCI and Lewy body immunoblotted with antibodies against total α-Syn or pS129 α-Syn (pSyn). **b**, Quantification of pS129 α-Syn versus total α-Syn shown in **a** (GCI, $n = 5$ cases; LB, $n = 7$ cases). **c**, Proteinase K-digested LB-α-Syn and GCI-α-Syn from six cases immunoblotted with anti-α-Syn MAb (Syn211). **d**, GCI-α-Syn and LB-α-Syn incubated with increasing concentrations of proteinase K (PK) and immunoblotted with Syn211 (experiment repeated three times). **e**, Semi-quantitative scores (0–3) to quantify α-Syn pathology revealed by Syn303 or Syn7015 immunohistochemistry in adjacent brain sections of patients with MSA or Lewy body disease (LB, $n = 9$ cases; GCI, $n = 7$ cases) (statistics: Mann–Whitney *U*-test). **f**, Quantification of area occupied by Syn7015-positive (Syn7015$^+$) versus Syn303-positive (Syn303$^+$) α-Syn pathology for experiments in **e** (LB, $n = 9$ cases, GCI, $n = 7$ cases). **g**, Representative photomicrographs for experiments in **e** (repeated with seven cases). **h**, Primary oligodendrocytes expressing α-Syn-mCherry incubated with 13 ng GCI-α-Syn, LB-α-Syn or PFFs were stained with 81A (pS129 α-Syn) and anti-olig2 (experiment repeated four times). **i**, Quantification of pS129 α-Syn induced by GCI-α-Syn, LB-α-Syn and PFF in oligodendrocytes expressing α-Syn (GCI, $n = 8$ different preparations; LB, $n = 9$ different preparations) (statistics: two-tailed unpaired *t*-test using the mean value of each case). **j**, Quantification of pS129 α-Syn induced by various amounts of PFFs, GCI-α-Syn or LB-α-Syn in oligodendrocytes expressing α-Syn ($n = 6$ (LB), 4 (GCI 3 ng) or 5 (all other groups) biological replicates) (statistics: adjusted with Bonferroni correction). Results shown as mean ± s.e.m. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$; NS, not significant. Scale bars: 100 µm (**g**, **h**), 25 µm (**g** inset), 10 µm (**h** inset). AD, Alzheimer's disease; DLB, dementia with Lewy bodies; LB, Lewy body; PDD, Parkinson's disease with dementia. For gel source data, see Supplementary Fig. 1. See Supplementary Table 5 for statistical details.

Data Fig. 4i, j), which suggests that the high potency of GCI-α-Syn is not likely to be due to its resistance to lysosomal proteolysis.

The markedly different seeding abilities of GCI-α-Syn and LB-α-Syn in vitro prompted us to test their properties in vivo. Equivalent amounts of GCI-α-Syn or LB-α-Syn were injected into the striatum of wild-type mice[20]. At three months post-injection, only GCI-α-Syn-injected wild-type (GCI-WT) mice developed abundant neuronal inclusions (Fig. 2g). At six months post-injection, although the amount of α-Syn pathology in GCI-WT mice declined markedly, only a limited number of neurons developed α-Syn pathology in LB-α-Syn-injected wild-type (LB-WT) mice (Extended Data Fig. 4k). Therefore, GCI-α-Syn more-potently induces neuronal pathology in vivo, which also argues against cell-type specific seeding by GCI-α-Syn. Furthermore, no oligodendroglial pathology was detected in GCI-WT mice, which further argues against the hypothesis that the properties of GCI-α-Syn dictate its oligodendrocyte distribution. On the other hand, the high potency of GCI-α-Syn probably contributes to the aggressive nature of MSA.

Misfolded α-Syn was shown to spread through the neuroanatomical connectome, but how the properties of different α-Syn strains affect this spreading process is unclear. We compared the transmission pattern of α-Syn pathology in wild-type mice injected with GCI-α-Syn, LB-α-Syn or PFFs. GCI-α-Syn spread more efficiently to the entorhinal cortex, hippocampus and pyramidal layer of piriform cortex, but less efficiently to other cortical regions such as the motor cortex, as compared to PFFs and LB-α-Syn (Fig. 2h, i and Extended Data Figs. 4l, 5a, b). By contrast, PFFs induced relatively more pathology in the striatum and substantia nigra but less in the olfactory bulb, as compared with LB-α-Syn and GCI-α-Syn (Fig. 2h, i and Extended Data Figs. 4l, 5a, b). Therefore, our data demonstrate that different α-Syn strains markedly modulate their transmission patterns in the nervous system.

The observation that the seeding properties of GCI-α-Syn and LB-α-Syn do not show any cell-type preference raises the question of why they demonstrate distinct cell-type distributions in diseased brains. Because our data argue against the hypothesis that strain properties determine cell-type distributions (hypothesis 1 in Extended Data Fig. 3g), we proposed an alternative hypothesis that the different cellular environments of neurons and oligodendrocytes promote the formation of distinct α-Syn strains (hypothesis 2 in Extended Data Fig. 3g). According to this hypothesis, injection of LB-α-Syn into mice expressing α-Syn in oligodendrocytes should convert the LB-α-Syn into a GCI-like strain. To test this hypothesis and to eliminate confounding factors arising from neuronal pathology, we crossed 2′,3′-cyclic nucleotide 3′ phosphodiesterase (CNP)-α-Syn transgenic mice (M2 line)[21] with α-Syn

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

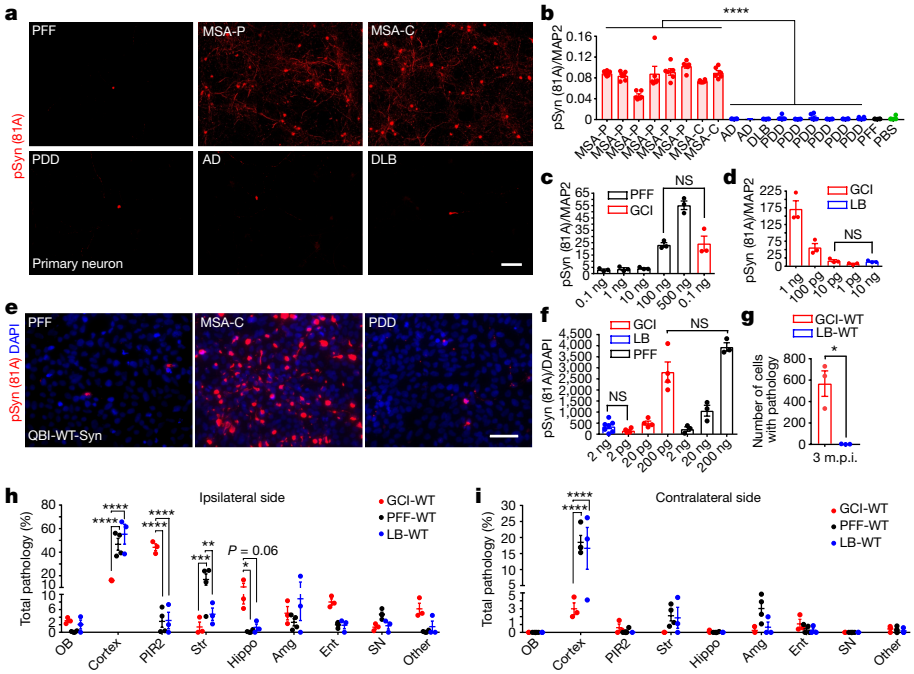


Fig. 2 | The seeding properties of GCI-α-Syn and LB-α-Syn do not show any cell-type preference. a, pS129 α-Syn induced by 1 ng of GCI-α-Syn, LB-α-Syn or PFFs in primary neurons (experiment repeated seven times). **b**, Quantification of α-Syn pathology for experiments in **a** (GCI, $n = 8$ different preparations, LB, $n = 9$ different preparations) (statistics: two-tailed unpaired t-test using the mean value of each case.). **c**, **d**, Quantification of pS129 α-Syn induced by various concentrations of PFFs, GCI-α-Syn or LB-α-Syn in neurons ($n = 3$ biological replicates) (statistics: adjusted with Bonferroni correction). **e**, pS129 α-Syn induced by 2 ng of GCI-α-Syn, LB-α-Syn or PFFs in QBI-WT-Syn cells (experiment repeated seven times). **f**, Quantification of pS129 α-Syn induced by various concentration of PFFs, GCI-α-Syn or LB-α-Syn in QBI-WT-Syn cells (LB, $n = 7$ (LB), 4 (GCI) or 3 (PFF) biological replicates) (statistics: adjusted with Bonferroni correction). **g**, Quantification of the number of cells with α-Syn pathology in wild-type mice inoculated with 50 ng of GCI-α-Syn or LB-α-Syn at three months post-injection (m.p.i.) ($n = 3$ mice). **h**, **i**, Quantification of the distribution of α-Syn pathology seeded by GCI-α-Syn (GCI-WT, $n = 3$ mice) or PFFs (PFF-WT, $n = 4$ mice) at three months post-injection and LB-α-Syn (LB-WT, $n = 3$ mice) at six months post-injection. Amg, amygdala; Cortex, cortex except pyramidal layer of piriform area (PIR2) and entorhinal cortex (Ent); Hippo, hippocampus; OB, olfactory bulb; SN, substantia nigra; Str, striatum. Results shown as mean ± s.e.m. Statistics shown in **h**, **i** are two-way ANOVA followed by Tukey's honest significant difference test. Scale bars: 100 μm (**a**), 50 μm (**e**). See Supplementary Table 5 for statistical details.

knockout mice, generating a new mouse line that expresses α-Syn only in oligodendrocytes (hereafter referred to as 'KOM2 mice') (Extended Data Fig. 6a, b).

KOM2 mice were unilaterally injected with GCI-α-Syn or LB-α-Syn into the thalamus (GCI-KOM2 and LB-KOM2 mice, respectively). At one month post-injection, although α-Syn inclusions appeared in the

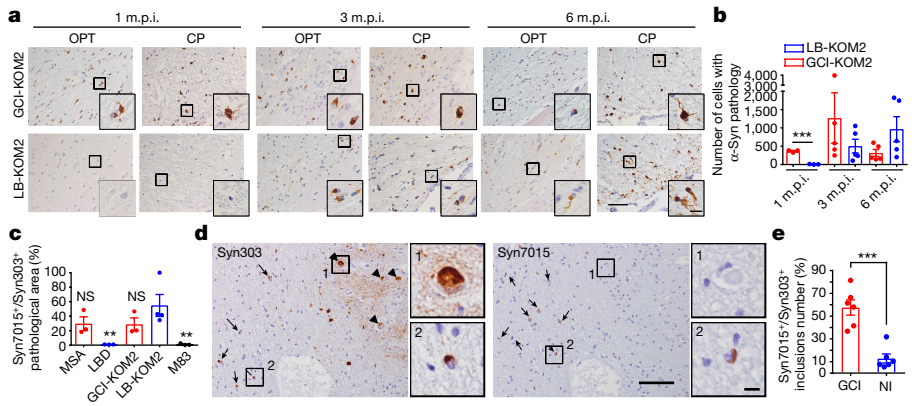


Fig. 3 | Oligodendrocyte environment generates the GCI-α-Syn strain. a, Syn506-positive α-Syn aggregates seeded by 18.75 ng of GCI-α-Syn or LB-α-Syn in KOM2 mice. OPT, optic tract; CP, cerebral peduncle (experiment repeated three times). **b**, Quantification of the number of oligodendrocytes with α-Syn pathology in injected KOM2 mice (one month post-injection, $n = 3$; three and six months post-injection, $n = 5$ mice) (statistics: adjusted with Bonferroni correction). **c**, Quantification of the ratio of area occupied by Syn7015-positive versus Syn303-positive α-Syn pathology in adjacent sections of brains from patients with MSA (MSA) or patients with Lewy body disease (LBD), injected KOM2 mouse brains or M83 transgenic mouse brains (GCI, $n = 3$ cases; LB, $n = 3$ cases; GCI-KOM2, LB-KOM2 and M83, $n = 4$ mice) (statistics: one-way ANOVA followed by Dunnett's post hoc test comparing each group with LB-KOM2). **d**, Adjacent sections of the medulla from an MSA case stained with Syn303 and Syn7015 (repeated with four cases). Arrows: GCIs; arrowheads: neuronal inclusions. **e**, Quantification of the ratio of Syn7015-positive versus Syn303-positive GCIs and neuronal inclusions (NI) on adjacent sections of brains from paients with MSA ($n = 6$ cases). Results shown as mean ± s.e.m. Scale bars: 50 μm (**a**); 12.5 μm (**a** insets); 100 μm (**d**); 10 μm (**d** insets). See Supplementary Table 5 for statistical details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

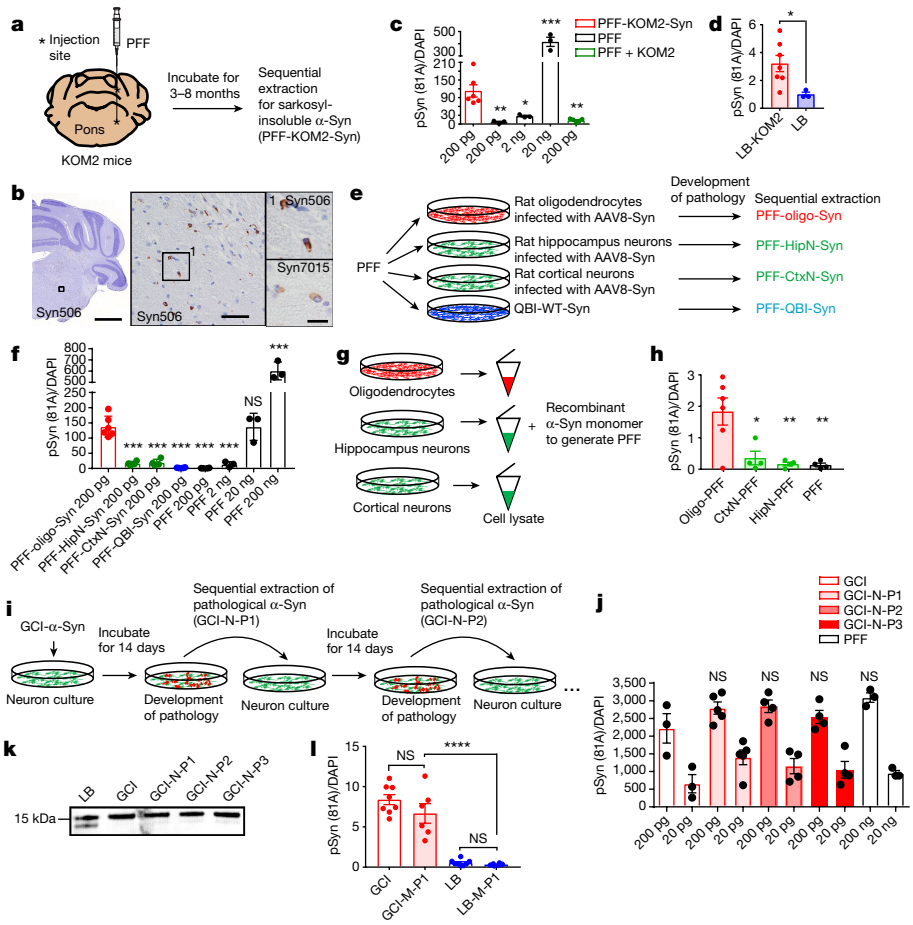


Fig. 4 | Oligodendrocytes convert misfolded α-Syn to a GCI-α-Syn-like strain but neurons could not convert GCI-α-Syn to a LB-α-Syn-like strain. a, Schematic for passaging PFFs in KOM2 mice. **b**, α-Syn pathology in PFF-injected KOM2 mice (experiment repeated three times). **c**, Quantification of pS129 α-Syn in QBI-WT-Syn cells seeded by PFFs, passaged PFFs (PFF-KOM2-Syn) or PFFs combined with insoluble fraction from uninjected KOM2 mice (PFF + KOM2) ($n = 6$ (PFF-KOM2-Syn), 3 (PFF) and 5 (PFF + KOM2) biological replicates). **d**, Quantification of pS129 α-Syn in QBI-WT-Syn cells seeded by equal amounts of LB-α-Syn or LB-α-Syn passaged in KOM2 mice (LB-KOM2) (LB, $n = 3$ biological replicates; LB-KOM2, $n = 7$ biological replicates). **e**, Schematic for passaging PFFs in various cells. **f**, Quantification of pS129 α-Syn in QBI-WT-Syn cells seeded by PFFs passaged in cells. PFF-oligo-Syn, PFF-HipN-Syn, PFF-CtxN-Syn and PFF-QBI-Syn refer to PFFs that have been passaged in oligodendrocytes, hippocampus neurons, cortical neurons and QBI-WT-Syn cells, respectively. ($n = 6$ (PFF-oligo-Syn), 4 (PFF-HipN-Syn, PFF-CtxN-Syn) and 3 (PFF) biological replicates). **g**, Schematic for generating α-Syn PFFs in cell lysates. **h**, Quantification of pS129 α-Syn in QBI-WT-Syn cells seeded by PFFs generated in oligodendrocyte lysate (oligo-PFF), cortex neuron lysate (CtxN-PFF), hippocampus neuron lysate (HipN-PFF) or with α-Syn monomer alone (PFF) ($n = 6$ (oligo-PFF) or 4 (CtxN-PFF, HipN-PFF, PFF) biological replicates). **i**, Schematic for passaging GCI-α-Syn in primary neurons. **j**, Quantification of pS129 α-Syn in QBI-WT-Syn cells seeded by α-Syn PFFs, GCI-α-Syn and GCI-α-Syn passaged in primary neurons ($n = 3$ (GCI, PFF), 4 (GCI-N-P2, GCI-N-P3) or 5 (GCI-N-P1) biological replicates). **k**, Proteinase K-digested LB-α-Syn, GCI-α-Syn, GCI-N-P1, GCI-N-P2 and GCI-N-P3 were immunoblotted with anti-α-Syn antibody (experiment repeated three times). **l**, Quantification of insoluble pS129 α-Syn in QBI-WT-Syn cells seeded by equal amounts of GCI-α-Syn, LB-α-Syn or GCI-α-Syn and LB-α-Syn passaged in M83 mice (GCI-M-P1 and LB-M-P1) ($n = 8$ (GCI, LB) or 6 (GCI-M-P1, LB-M-P1) biological replicates). Results shown as mean ± s.e.m. Statistics shown in **c**, **f**, **h** and **j** are one-way ANOVA followed by Dunnett's post hoc test comparing each group with PFF-KOM2-Syn (**c**), PFF-oligo-Syn (**f**), oligo-PFF (**h**) or GCI 200 pg (**j**). Statistics shown in **l** are one-way ANOVA followed by Tukey's multiple comparison's test. Scale bars: 1 mm (**b**); 50 μm (**b**, middle inset), 15 μm (**b**, right inset). For gel source data, see Supplementary Fig. 1. See Supplementary Table 5 for statistical details.

thalamus and fimbria of GCI-KOM2 mice, more abundant pathology was observed in the optic tract and cerebral peduncle; that is, sites distant from the injection site but highly enriched with oligodendrocytes, which suggests that α-Syn pathology spreads between oligodendrocytes. By contrast, few oligodendrocytes developed pathology in LB-KOM2 mice, demonstrating that GCI-α-Syn is more potent than LB-α-Syn. However, at three months post-injection—as the burden of pathology in GCI-KOM2 mice peaked—LB-KOM2 mice began to show substantial oligodendrocyte pathology. At six months post-injection, whereas the pathology in GCI-KOM2 mice declined precipitously, LB-KOM2 mice showed even more inclusions, which reached levels comparable to GCI-KOM2 mice at three months post-injection (Fig. 3a, b and Extended Data Fig. 7a–e). The presence of α-Syn pathology in oligodendrocytes and the phosphorylation of S129 were confirmed by immunostaining (Extended Data Fig. 7f, g). Mapping of α-Syn pathology in LB-KOM2 and GCI-KOM2 mice revealed a similar pattern (Extended Data Fig. 8). This delayed induction of α-Syn pathology in LB-KOM2 mice supports our hypothesis that LB-α-Syn is less potent than GCI-α-Syn, but that once initiated the subsequent propagation within oligodendrocytes results in the formation of a GCI-α-Syn-like strain.

Moreover, whereas Lewy bodies in the brains of patients with Parkinson's disease were not detected by Syn7015, the oligodendrocyte α-Syn pathologies induced by LB-α-Syn in KOM2 mice were Syn7015-positive (Fig. 3c and Extended Data Fig. 9a), suggesting that LB-α-Syn-induced oligodendrocyte pathology acquired properties of the GCI-α-Syn strain. Furthermore, detailed characterization of oligodendrocyte pathologies induced by GCI-α-Syn or LB-α-Syn

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

showed that they were indistinguishable, including co-localization with p62, partial co-localization with ubiquitin and association with reactive astrocytes (Extended Data Fig. 9b, c). Thus, LB-α-Syn can induce oligodendrocyte pathologies in the KOM2 mice that adopt properties of the GCI-α-Syn strain.

Both α-Syn neuronal inclusions and GCIs are present in the brains of patients with MSA, although neuronal inclusions are uncommon. We hypothesized that if the oligodendrocyte environment generates the GCI-α-Syn strain, neuronal inclusions in brains from paients with MSA would resemble Lewy bodies in brains from patients with Parkinson's disease. To test this, adjacent sections from six cases of MSA with neuronal inclusions in the medulla or hippocampus were immunostained with Syn7015 and Syn303. Syn7015 preferentially stained GCIs over neuronal inclusions (Fig. 3d, e), demonstrating that MSA neuronal inclusions are similar to Parkinson's disease Lewy bodies. Furthermore, Syn7015 also failed to detect neuronal inclusions in other brain regions—such as the substantia nigra and cortex—in additional cases of MSA (Extended Data Fig. 9d), which also provides support for our hypothesis.

To test our hypothesis further, we injected human α-Syn PFFs into the pons and cerebellum of KOM2 mice (Fig. 4a, b). Then, the induced pathological α-Syn (PFF-KOM2-Syn) was recovered by sequential extraction. Notably, PFF-KOM2-Syn was much more potent than the PFFs themselves in inducing α-Syn pathology (Fig. 4c and Extended Data Fig. 10a). By contrast, mixing PFFs with the sarkosyl-insoluble fraction from uninjected KOM2 mice only slightly increased the potency (Fig. 4c). A similar phenomenon was observed when passaging LB-α-Syn in KOM2 mice (Fig. 4d). To test whether this effect of the cellular milieu is unique to oligodendrocytes, PFFs were added to different cell types including oligodendrocytes, hippocampal and cortical neurons, and QBI-WT-Syn cells. After the induction of α-Syn pathology, sarkosyl-insoluble α-Syn was prepared from these cells (Fig. 4e). PFFs passaged through oligodendrocytes were more potent than PFFs passaged through other cell types (Fig. 4f and Extended Data Fig. 10b–d). Taken together, these results clearly demonstrate that the oligodendrocyte environment leads to the generation of the GCI-α-Syn strain.

To test whether the generation of the GCI-α-Syn strain depends on cell structures or specific 'factors', we prepared cell lysates from primary oligodendrocytes and neurons in which cell structures were disrupted but cell factors were preserved. We incubated each lysate with an equal amount of α-Syn monomer to generate α-Syn PFFs (Fig. 4g). PFFs generated in oligodendrocyte lysates were able to induce several-fold more pathology compared with PFFs generated in neuron lysates or with α-Syn monomers alone, supporting the hypothesis that the generation of the GCI-α-Syn strain relies on specific factors in oligodendrocytes (Fig. 4h).

Lastly, we asked whether the neuronal environment could convert the GCI-α-Syn strain to the LB-α-Syn strain. Primary mouse neurons were treated with GCI-α-Syn, and the induced α-Syn pathology was enriched by sequential extraction (pathological α-Syn labelled GCI-N-P1, GCI-N-P2 and GCI-N-P3 over three rounds of passaging) (Fig. 4i). Using enzyme-linked immunosorbent assays (ELISAs) that detect only human α-Syn or both human and mouse α-Syn, we estimated that majority (>99.7%) of pathological α-Syn in GCI-N-P1 was derived through recruitment of mouse α-Syn (Supplementary Table 3). GCI-N-P1 was as potent as GCI-α-Syn (Fig. 4j). Furthermore, the GCI-α-Syn strain was repetitively passaged in neurons such that even after three rounds of passaging GCI-N-P3 still maintained the marked activity of the GCI-α-Syn strain (Fig. 4j and Extended Data Fig. 10e–g). Proteinase K digestion revealed that after passaging in neurons, the GCI-α-Syn conformation was maintained (Fig. 4k). Furthermore, when both GCI-α-Syn and LB-α-Syn were passaged in M83 mice expressing mutant human (A53T) α-Syn[22], their different seeding properties were maintained (Fig. 4l). Thus, we conclude that the GCI-α-Syn strain can be maintained in a neuronal milieu. Combined with the observation that neuronal inclusions in MSA are similar to LB-α-Syn, our data suggest that GCI-α-Syn rarely transmits from oligodendrocytes to neurons in MSA brains, although we cannot exclude the possibility that GCI-α-Syn transmits to neurons, but that these neurons died.

Because LB-α-Syn could induce α-Syn pathology in oligodendrocytes expressing α-Syn in vitro and in vivo, the lack of GCIs in Lewy body diseases is unlikely to be due to the strain properties of LB-α-Syn, but might be attributable to the lack of α-Syn in oligodendrocytes[23]. The source of α-Syn in oligodendrocytes in MSA is still unclear. Two hypotheses have previously been proposed: (1) oligodendrocytes pathologically overexpress α-Syn in MSA[24] and (2) oligodendrocytes take up α-Syn from neurons[25]. GCI-α-Syn also could not induce α-Syn pathology in oligodendrocytes cultured in medium with high concentration of α-Syn monomer (data not shown), suggesting that internalization of exogenous α-Syn from the environment might be insufficient for GCI formation in oligodendrocytes.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0104-4.

Received: 2 August 2017; Accepted: 6 April 2018;
Published online: 09 May 2018

1. Lippa, C. F. et al. Lewy bodies contain altered α-synuclein in brains of many familial Alzheimer's disease patients with mutations in presenilin and amyloid precursor protein genes. *Am. J. Pathol.* **153**, 1365–1370 (1998).
2. Spillantini, M. G. et al. α-Synuclein in Lewy bodies. *Nature* **388**, 839–840 (1997).
3. Tu, P. H. et al. Glial cytoplasmic inclusions in white matter oligodendrocytes of multiple system atrophy brains contain insoluble α-synuclein. *Ann. Neurol.* **44**, 415–422 (1998).
4. Bousset, L. et al. Structural and functional characterization of two α-synuclein strains. *Nat. Commun.* **4**, 2575 (2013).
5. Guo, J. L. et al. Distinct α-synuclein strains differentially promote tau inclusions in neurons. *Cell* **154**, 103–117 (2013).
6. Peelaerts, W. et al. α-Synuclein strains cause distinct synucleinopathies after local and systemic administration. *Nature* **522**, 340–344 (2015).
7. Prusiner, S. B. et al. Evidence for α-synuclein prions causing multiple system atrophy in humans with parkinsonism. *Proc. Natl Acad. Sci. USA* **112**, E5308–E5317 (2015).
8. Woerman, A. L. et al. Propagation of prions causing synucleinopathies in cultured cells. *Proc. Natl Acad. Sci. USA* **112**, E4949–E4958 (2015).
9. Fanciulli, A. & Wenning, G. K. Multiple-system atrophy. *N. Engl. J. Med.* **372**, 249–263 (2015).
10. Koga, S. & Dickson, D. W. Recent advances in neuropathology, biomarkers and therapeutic approach of multiple system atrophy. *J. Neurol. Neurosurg. Psychiatry* **89**, 175–184 (2018).
11. Fujiwara, H. et al. α-Synuclein is phosphorylated in synucleinopathy lesions. *Nat. Cell Biol.* **4**, 160–164 (2002).
12. Gorbatyuk, O. S. et al. The phosphorylation state of Ser-129 in human α-synuclein determines neurodegeneration in a rat model of Parkinson disease. *Proc. Natl Acad. Sci. USA* **105**, 763–768 (2008).
13. Covell, D. J. et al. Novel conformation-selective alpha-synuclein antibodies raised against different *in vitro* fibril forms show distinct patterns of Lewy pathology in Parkinson's disease. *Neuropathol. Appl. Neurobiol.* **43**, 604–620 (2017).
14. Duda, J. E., Giasson, B. I., Mabon, M. E., Lee, V. M. & Trojanowski, J. Q. Novel antibodies to synuclein show abundant striatal pathology in Lewy body diseases. *Ann. Neurol.* **52**, 205–210 (2002).
15. Irwin, D. J. et al. Neuropathologic substrates of Parkinson disease dementia. *Ann. Neurol.* **72**, 587–598 (2012).
16. Montine, T. J. et al. Multisite assessment of NIA-AA guidelines for the neuropathologic evaluation of Alzheimer's disease. *Alzheimers Dement.* **12**, 164–169 (2016).
17. Volpicelli-Daley, L. A. et al. Exogenous α-synuclein fibrils induce Lewy body pathology leading to synaptic dysfunction and neuron death. *Neuron* **72**, 57–71 (2011).
18. Luk, K. C. et al. Exogenous α-synuclein fibrils seed the formation of Lewy body-like intracellular inclusions in cultured cells. *Proc. Natl Acad. Sci. USA* **106**, 20051–20056 (2009).
19. Karpowicz, R. J. Jr et al. Selective imaging of internalized proteopathic α-synuclein seeds in primary neurons reveals mechanistic insight into transmission of synucleinopathies. *J. Biol. Chem.* **292**, 13482–13497 (2017).
20. Luk, K. C. et al. Pathological α-synuclein transmission initiates Parkinson-like neurodegeneration in nontransgenic mice. *Science* **338**, 949–953 (2012).
21. Yazawa, I. et al. Mouse model of multiple system atrophy α-synuclein expression in oligodendrocytes causes glial and neuronal degeneration. *Neuron* **45**, 847–859 (2005).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

22. Giasson, B. I. et al. Neuronal α-synucleinopathy with severe movement disorder in mice expressing A53T human α-synuclein. *Neuron* **34**, 521–533 (2002).
23. Miller, D. W. et al. Absence of α-synuclein mRNA expression in normal and multiple system atrophy oligodendroglia. *J. Neural Transm. (Vienna)* **112**, 1613–1624 (2005).
24. Asi, Y. T. et al. Alpha-synuclein mRNA expression in oligodendrocytes in MSA. *Glia* **62**, 964–970 (2014).
25. Reyes, J. F. et al. Alpha-synuclein transfers from neurons to oligodendrocytes. *Glia* **62**, 387–398 (2014).

Acknowledgements We thank K. Brunden and J. Guo for critical reading of the manuscript; S. Xie for helping with statistics; L. Kwong and Y. Xu for helping with brain extraction; D. Riddle for providing primary neurons; L. Romero for helping with quantification; L. Changolkar for helping with ELISA and all the other members of the Center for Neurodegenerative Disease Research for their support; and J. Grinspan and C. Richter-Landsberg for advice on culturing oligodendrocytes. The anti-PLP antibody was provided by J. Grinspan. This work was supported by NIH/NINDS Udall Center grant NS53488, the Ofer Nimerovsky Family Fund, the Jeff and Anne Keefer Fund and the MSA Coalition Global Seed Grant.

Reviewer information *Nature* thanks T. Baron and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.P. designed and performed the experiments, analysed the data and wrote the manuscript. R.J.G., C.M. and M.F.O. performed the animal experiments and did quantification analysis. R.J.G. also performed α-Syn ELISA and biochemical analysis. D.J.C. generated the Syn9027 and Syn7015 antibodies and α-Syn PFFs. A.S. generated the KOM2 mice and assisted in stereotaxic injections. J.L.R. scored stained human brain sections. B.Z., R.J.G. and M.F.O. performed stereotaxic injections. R.M.P. performed cell culture experiments. K.C.L performed experiments and reviewed the manuscript. J.Q.T. supervised the study and reviewed the manuscript. V.M.-Y.L. supervised the study, interpreted the data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0104-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0104-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to V.M.-Y.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The investigators were blinded to allocation during all quantification experiments involving manual counting.

Recombinant α-Syn purification and in vitro fibrillization. Full-length human and mouse α-Syn (1–140) proteins were expressed in BL21 (DE3) RIL cells and purified as previously described[26]. Fibrillization was conducted by diluting recombinant α-syn to 5 mg/ml in sterile Dulbecco's PBS (Cellgro, Mediatech; pH adjusted to 7.0) followed by incubating this recombinant α-Syn at 37 °C with constant agitation at 1,000 r.p.m. for 7 days. Successful α-Syn fibrillization was verified by sedimentation test and thioflavin T-binding assay as described[27].

Preparation of sarkosyl-insoluble fractions from disease and control brains. All human brain tissues are from the CNDR brain bank[28]. Preparations of the sarkosyl-insoluble fraction were performed as previously described[5] and outlined in Extended Data Fig. 1a. In brief, brain regions with abundant α-Syn inclusions from patients with MSA, PDD, DLB or AD were identified by post mortem histological examination[15]. The brains of patients with Alzheimer's disease were selected for the presence of abundant Lewy bodies in addition to plaques and tangles. Frozen brain tissues from the identified regions were homogenized in high-salt (HS) buffer (50 mM Tris-HCl pH 7.4, 750 mM NaCl, 10 mM NaF, 5 mM EDTA) with protease and protein phosphatase inhibitors, incubated on ice for 20 min and centrifuged at 100,000 *g* for 30 min. The pellets were then re-extracted with HS buffer, followed by sequential extractions with 1% Triton X-100-containing HS buffer and 1% Triton X-100-containing HS buffer with 30% sucrose. The pellets were then re-suspended and homogenized in 1% sarkosyl-containing HS buffer, rotated at 4 °C overnight and centrifuged at 100,000 *g* for 30 min. The resulting sarkosyl-insoluble pellets were washed once with Dulbecco's PBS and re-suspended in Dulbecco's PBS by brief sonication (QSonica Microson XL-2000; 20 pulses; setting 2; 0.5 s per pulse). This suspension was termed the 'sarkosyl-insoluble fraction', which contained pathological α-Syn and was used for the cellular and in vivo assays described in this Letter. The amount of α-Syn, tau, Aβ 40 and Aβ 42 in the sarkosyl-insoluble fractions was determined by sandwich ELISAs (see 'Sandwich ELISA') and the protein concentrations were examined by bicinchoninic acid assay. Proteinase K, trypsin and thermolysin digestion was performed as previously described[5].

Sandwich ELISA. The concentrations of tau, Aβ 1–40 and Aβ 1–42 in the sarkosyl-insoluble fraction of human brain extractions were measured using sandwich ELISA as previously described[29,30], with the following combinations of capture and reporting antibodies: Tau5/BT2^{+}HT7 for tau, Ban50/BA27 for Aβ 1–40, Ban50/BC05 for Aβ 1–42.

To measure the concentration of α-Syn, 384-well Nunc Maxisorp clear plates were coated with 100 ng (30 μl per well) Syn9027, a MAb to α-Syn, in Takeda buffer and incubated overnight at 4 °C. The plates were washed 4× with PBS containing 1% Tween 20 (PBS-T), and blocked using Block Ace solution (90 μl per well) (AbD Serotec) overnight at 4 °C. Then, the plates were incubated with brain lysates at 4 °C overnight using recombinant α-Syn monomer as standards. The plates were then washed with PBS-T and a rabbit monoclonal anti-α-Syn antibody, MJF-R1 (1:1,000, 30 μl per well) was added to each well and incubated at 4 °C overnight. After washing, goat-anti-rabbit-IgG conjugated to horse radish peroxidase (Cell Signaling Technology, 1:15,000, 30 μl per well) was added to the plates followed by incubation for 2 h at room temperature. Following another wash, the plates were developed for 10–15 min using 1-Step Ultra TMB-ELISA substrate solution (Fisher Scientific, 30 μl per well), the reaction was quenched using 10% phosphoric acid (30 μl per well) and plates were read at 450 nm on a Molecular Devices Spectramax M5 plate reader.

Cell cultures. Primary mouse neurons were prepared from the hippocampus of embryonic day (E) 15–E17 CD1 mouse embryos as previously described[27]. PFFs and sarkosyl-insoluble α-Syn fractions were diluted in Dulbecco's PBS (without Mg^{2+} or Ca^{2+}) and sonicated (QSonica Microson XL-2000; 60 pulses; setting 1.5; 0.5 s per pulse). Neurons were then treated with PBS, sonicated PFFs or the α-Syn sarkosyl-insoluble fractions at 10 days in vitro (DIV) and collected for immunocytochemistry at 14 days post-treatment. To passage GCIs in primary neurons, five million neurons were plated per 10 cm dish, treated with 30 ng of sonicated GCI-α-Syn at DIV 10 and collected at DIV 24 by sequential extraction with the same protocol as described for human brain, except that HS buffer (50 mM Tris-HCl pH 7.4, 750 mM NaCl, 10 mM NaF, 5 mM EDTA) containing 1% Triton-100 was used in the initial extraction. The amount of total α-Syn in the sarkosyl-insoluble fractions was determined by sandwich ELISA with antibodies against both mouse and human α-Syn (Syn9027 and HuA) and the amount of human α-Syn was examined by sandwich ELISA using MJF-R1, which recognizes only human α-Syn, as the reporter antibody. Proteinase K digestion was performed as above described for GCI-α-Syn and LB-α-Syn. Treatment of primary neurons with chloroquine was performed as previously described[19].

Primary oligodendrocytes were prepared from the cortex of neonatal Sprague Dawley rats (Charles River Laboratories) as previously described[31]. In brief, oligodendrocyte progenitor cells were purified from mixed glial culture by mechanical shaking. The purified oligodendrocyte progenitor cells were plated on poly-L-lysine-coated coverslips and infected with AAV8-α-Syn-mCherry or AAV8-α-Syn three days after plating. Then, differentiation was induced three days after infection by culturing oligodendrocyte progenitor cells in the differentiation medium. Treatment with sonicated α-Syn PFF and sarkosyl-insoluble fractions was performed three days after differentiation as described above. Treated oligodendrocytes were collected for immunocytochemistry at 14 days post-treatment. To evaluate the purity of the oligodendrocyte culture, three independent cultures were stained with various cell markers for astrocytes (glial fibrillary acidic protein, GFAP), microglial (Iba1), neuron (NeuN) and oligodendrocytes (Olig2) at DIV 3, DIV 9 and DIV 23. At least three different 20× images were randomly taken for each coverslip and at least three coverslips were analysed for each time point for each cell marker.

To passage PFF in primary oligodendrocytes, two million oligodendrocyte progenitor cells were plated in a 10-cm dish, infected with AAV8-Syn to express human α-Syn, differentiated and treated with sonicated α-Syn PFFs as described above. Cells were collected at 14 days after treatment by sequential extraction with the same protocol as described for primary mouse neurons. The amount of total α-Syn in the sarkosyl-insoluble fractions was determined by sandwich ELISA with three different combinations of capture and reporting antibodies (Syn9027 + HuA, Syn9027 + MJF-R1 and HuA + Syn211).

The culture of QBI-WT-Syn cells and treatment with misfolded α-Syn were performed as previously described[18]. To induce α-Syn pathology in QBI-WT-Syn cells, one million QBI-WT-Syn cells were plated in a 6-cm dish and treated with PFFs two days later. Cells were collected at three days post-treatment by sequential extraction with the same protocol as described for primary mouse neurons. The amount of total α-Syn in the sarkosyl-insoluble fractions was determined by sandwich ELISA with three different combinations of capture and reporting antibodies (Syn9027 + HuA, Syn9027 + MJF-R1 and HuA + Syn211).

Primary hippocampus and cortical neurons of Sprague Dawley rats were generated from the Neuron Culture Service Center at University of Pennsylvania. To passage PFFs in primary rat neurons, 1.2 million hippocampus or cortical neurons were plated in a 6-cm dish, infected with AAV8-α-Syn at DIV 3 to express human α-Syn and treated with PFFs at DIV 6 as described above. Cells were collected at 14 days post-treatment by sequential extraction with the same protocol as described for primary mouse neurons. The amount of total α-Syn in the sarkosyl-insoluble fractions was determined by sandwich ELISA with three different combinations of capture and reporting antibodies (Syn9027 + HuA, Syn9027 + MJF-R1 and HuA + Syn211).

Stereotaxic injection of sarkosyl-insoluble fraction of pathological α-Syn and α-Syn PFFs. Sarkosyl-insoluble fractions from diseased brains were diluted in sterile Dulbecco's PBS to reach the same concentration of α-Syn in the samples and sonicated as described above. Two-to-three-month-old C57BL6/C3H wild-type mice or 3–4-month-old KOM2 mice were anaesthetized with ketamine hydrochloride (100 mg/kg), xylazine (10 mg/kg) and acepromazine (0.1 mg/kg). For wild-type mice, 50 ng of sarkosyl-insoluble pathological α-Syn from two different MSA brains and one brain of a patient with PDD ('PDD brain') or 6.25 μg mouse PFFs in 2.5 μl Dulbecco's PBS was stereotaxically injected into the dorsal striatum (coordinates: 0.2 mm relative to bregma, +2.0 mm from midline, +3.2 mm beneath the surface of skull) with 10-μl syringes (Hamilton) at a rate of 0.4 μl per min. For KOM2 mice, 18.75 ng of pathological α-Syn from three different MSA brains, two different PDD brains and one brain of a patient with DLB ('DLB brain') in 2.5 μl Dulbecco's PBS was stereotaxically injected into the thalamus, an area of relatively high oligodendrocyte α-Syn expression in these mice (coordinates: −2.5 mm relative to bregma, +2.0 mm from midline, +3.4 mm beneath the surface of skull) at a rate of 0.4 μl per min. Animals were then euthanized at 1, 3 and 6 month post-injection.

To passage PFFs in KOM2 mice, 5 mg/ml PFFs were sonicated as described above and stereotaxically injected into the pons and cerebellum of the KOM2 mice (coordinates: −5.45 mm relative to bregma, +1.1 mm from midline, +5 mm beneath the surface of skull for pons and +2.6 mm for cerebellum). PFFs were injected into pons (1 μl of PFFs) and cerebellum (1.5 μl) at a rate of 0.1 μl per min. Mice were euthanized at 3–8 months post-injection and the pons and cerebellum were either processed for histological studies or frozen for sequential extraction. Sequential extractions were performed with the same protocol as described for human brain except that the two rounds of extraction of HS buffer were omitted and the extraction began with 1% Triton, containing HS buffer. The amount of pathological α-Syn in the sarkosyl-insoluble fraction was examined by sandwich ELISAs with two different combinations of capture and reporting antibodies (9027 + MJF-R1 and Syn211 + HuA). To passage LB-α-Syn in KOM2, 2.5 μl of LB-α-Syn at the concentration of 7.52–15.87 ng/μl was injected into the pons and cerebellum or the thalamus of KOM2 mice. Mice were euthanized at 3–8 months post-injection and the brains were sequentially extracted and analysed in the same

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

way as PFF-injected KOM2 mice. To passage LB-α-Syn and GCI-α-Syn in M83 mice, 2.5 μl of GCI-α-Syn or LB-α-Syn at the concentration of 7.52 ng/μl was injected bilaterally into the striatum of M83 mice. Mice were euthanized at one month post-injection and the brains were sequentially extracted and analysed in the same way as PFF-injected KOM2 mice.

Immunohistochemistry. For histological studies, animals were transcardially perfused with PBS and the brain and spinal cord were removed and fixed in 70% ethanol (in 150 mM NaCl, pH 7.4) overnight before being processed for paraffin embedding. Immunohistochemistry was performed on 6-μm thick sections as previously described[32,33]. The names and dilutions of the primary antibodies used here are shown in Supplementary Table 4. Stained sections were digitized using a Perkin Elmer Lamina scanner at 20× magnification. For quantitative analysis of α-Syn pathology in the mouse brain, Syn506-stained sections spanning the entire mouse brain at ~120-μm intervals were counted manually for the total number of cells with α-Syn positive inclusions. The semiquantitative heat maps were generated as previously described[34] using Syn506-stained slides. To quantify the spreading of pathological α-Syn in wild-type mice, representative Syn506-stained brain sections at bregma 4.28, 2.10, 0.98, −0.22, −1.22, −2.18, −2.92, −3.52 and −4.48 mm were counted manually for the number of cells with α-Syn pathology in each brain region. For GCI-WT and PFF-WT mice, 1–3 brain sections at each bregma point have been quantified and the mean value of these 1–3 brain sections has been used for quantification. For LB-WT mice, because of the low amount of pathology, 3–13 brain sections at each bregma point have been quantified.

For the quantification of neuronal inclusions in cases of MSA, adjacent sections of medulla or hippocampus from six cases of MSA were stained with Syn7015 and Syn303. The stained sections were digitized and the number of neuronal inclusions labelled by each antibody was quantified manually for each inferior olivary nucleus in medulla and dentate gyrus in hippocampus. To quantify the number of stained GCIs, three to six 10× images were randomly sampled in the white matter region near each inferior olivary nucleus or dentate gyrus. The positions of these 10× images were digitally labelled by drawing lines on the boundaries, and the same 10× images were sampled on adjacent sections by copying these annotations using the digitized image, and the total number of GCIs labelled by Syn7015 or Syn303 in these sampled areas were counted manually.

To grade the α-Syn pathology revealed by Syn7015 or Syn303 immunohistochemistry in MSA and LB-spectrum α-synucleinopathies, adjacent sections of hippocampus, frontal cortex, substantia nigra, cerebellum or midbrain from three cases of AD, three cases of DLB, three cases of PDD, four cases of MSA-C and three cases of MSA-P (see Supplementary Table 1) were stained with Syn7015 and Syn303 and graded by experienced pathologists. To quantify the ratio of total amount of Syn7015-positive versus Syn303-positive GCI or Lewy body, the stained adjacent sections were digitized and the total area occupied by Syn7015-positive or Syn303-positive α-Syn inclusions was quantified using an automated threshold-based algorithm by HALO software (http://www.indicalab.com/halo/). Then, the ratios of total Syn7015-positive area versus Syn303-positive area were calculated. For the serial dilution of Syn7015 and Syn303, all the ratios were calculated against the area occupied by immunohistochemistry conducted with the highest concentration of Syn303 (45 ng ml).

Animals. Two-to-three-month-old C57BL6/C3H wild-type mice were purchased from the Jackson Laboratories (Bar Harbour). M2 mice expressing wild-type human α-Syn in oligodendrocytes under the control of CNP promoter have previously been described[21]. KOM2 mice were generated by crossing M2 mice with $Snca^{-/-}$ mice[35]. Three-to-four-month-old KOM2 mice have been used for the study. All breeding, housing, and experimental procedures were performed according to the NIH Guide for the Care and Use of Experimental Animals and approved by the University of Pennsylvania Institutional Animal Care and Use Committee (IACUC). Both male and female mice were used for this study. Mice were randomly assigned to each experimental group. Roughly equal numbers of male and female KOM2 mice have been used in each group. All the GCI-WT and LB-WT mice were female. All the three-months-post-injection PFF-WT mice were female. For the six-months-post-injection PFF-WT mice, three of them were male and one was female.

Immunocytochemistry and quantification. For regular immunocytochemistry, cells were fixed in 4% paraformaldehyde (PFA) for 15 min followed by permeabilization with 0.1% Triton X-100 for 15 min. To examine α-Syn aggregates, cells were fixed with 4% PFA containing 1% Triton X-100 for 15 min to remove soluble proteins. Fixed coverslips were blocked with 3% BSA and 3% FBS for 1 h at room temperature and incubated with specific primary antibodies (Supplementary Table 4) at 4 °C overnight followed by staining with secondary antibodies for 2 h at room temperature. After mounting with Fluoromount G with DAPI (eBioscience), coverslips were scanned on a Perkin Elmer Lamina scanner. The total amount of 81A signal, the total amount of MAP2 signals for neuronal cultures and the total number of DAPI-positive nuclei for oligodendrocytes and QBI-WT-Syn cells were quantified using Indica Labs HALO software. For cells cultured in 96-well plates, cells were incubated in DAPI solution after staining with secondary antibodies. Then, plates were scanned with In Cell Analyzer 2200 (GE Healthcare) and analysed using the accompanying software (In Cell Toolbox Analyzer).

Purification and depletion of α-Syn from the sarkosyl-insoluble fraction by immunoprecipitation. Control mouse IgG (Sigma) or Syn9027 MAb—an in-house-generated Mab against α-Syn (epitope aa130–140)—were coupled to tosyl-activated Dynabeads (Invitrogen) or NHS-activated magnetic beads (Thermo Scientific) following the manufacturer's instructions. For immunoprecipitation purification, sarkosyl-insoluble fractions from diseased brains were incubated with control IgG-coupled beads in Dulbecco's PBS and rotated at 4 °C overnight. The resulting supernatant was then incubated with Syn9027-coupled beads in a rotator at 4 °C overnight to capture α-Syn. The following day, the Syn9027 beads were washed 3 times with Dulbecco's PBS and incubated with 0.1 M ethanolamine (pH 11.5) for 3–7 min at 55 °C to elute the bound α-Syn, which was then neutralized immediately with 1 M Tris (pH 7.0) and the eluted samples were stored at −80 °C until use. For immunoprecipitation depletion, the sarkosyl-insoluble fractions from diseased brains were incubated with Syn9027-coupled beads at 4 °C overnight. The resulting supernatants were incubated with Syn9027 beads again for a second round of α-Syn depletion and the final supernatants were stored at −80 °C until use.

Evaluatuing α-Syn expression in KOM2 mice. Total proteins were extracted from KOM2 and wild-type mice by sonicating the brain in 1% Triton X-100-containing HS buffer with phosphatase and protease inhibitors, and centrifuging at 100,000*g* for 30 min at 4 °C. Protein concentration of the supernatant was determined by BCA assay and same amount of protein from each mouse was resolved on 12% Bis-Tris gel and immunoblotted with antibodies against total α-Syn, mouse α-Syn or β-tubulins.

Generation of PFFs in different cell lysates. Neuron and oligodendrocyte cell lysates were prepared by sonicating primary rat hippocampus or cortical neurons at DIV 12 and oligodendrocyte cultures at DIV 10–12 in Dulbecco's PBS. The protein concentrations of cell lysates were evaluated by BCA assay and adjust to 1.86 mg/ml. α-Syn monomer was added to these cell lysates at the final concentration of 500 μg/ml and shaken at 37 °C with constant agitation at 1,000 r.p.m. for 14 days.

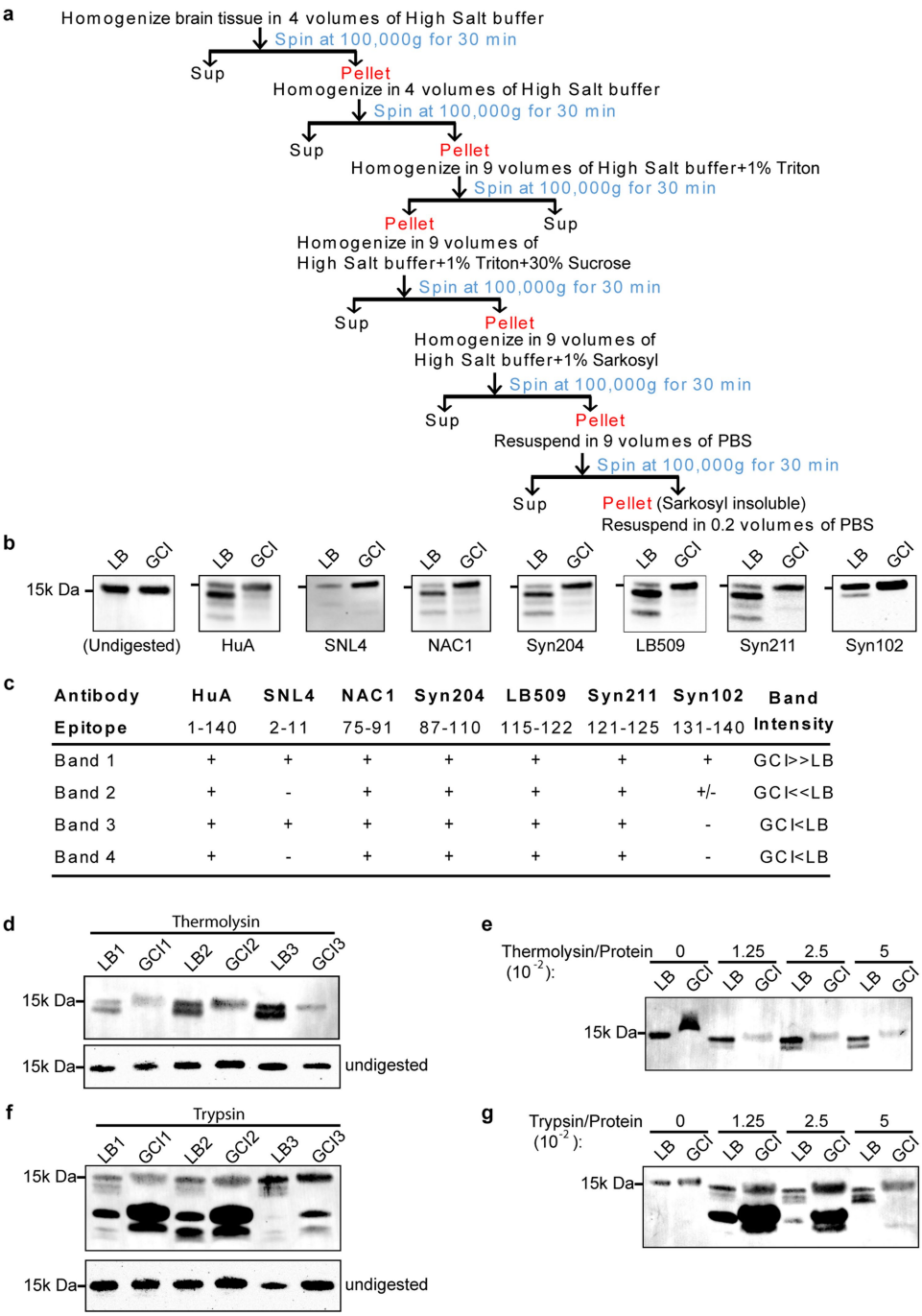
Statistical analysis. Unless specified otherwise, a two-tailed unpaired Student's *t*-test was used for all the comparisons in the study, and differences with *P* values of less than 0.05 were considered significant. For each *t*-test, an *F* test was also performed to evaluate the differences in variances. If there was a significant difference in variances ($P < 0.05$ by *F* test), Welch's correction on *t*-test was performed. Multiple comparisons were adjusted with Bonferroni correction. Detailed information regarding statistical analyses is provided in Supplementary Table 5.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All data generated or analysed during this study are included in the Letter and its Supplementary Information.

26. Giasson, B. I., Murray, I. V., Trojanowski, J. Q. & Lee, V. M. A hydrophobic stretch of 12 amino acid residues in the middle of α-synuclein is essential for filament assembly. *J. Biol. Chem.* **276**, 2380–2386 (2001).
27. Volpicelli-Daley, L. A., Luk, K. C. & Lee, V. M. Addition of exogenous α-synuclein preformed fibrils to primary neuronal cultures to seed recruitment of endogenous α-synuclein to Lewy body and Lewy neurite-like aggregates. *Nat. Protocols* **9**, 2135–2146 (2014).
28. Toledo, J. B. et al. A platform for discovery: the University of Pennsylvania Integrated Neurodegenerative Disease Biobank. *Alzheimers Dement.* **10**, 477–484.e1 (2014).
29. Guo, J. L. & Lee, V. M. Seeding of normal Tau by pathological Tau conformers drives pathogenesis of Alzheimer-like tangles. *J. Biol. Chem.* **286**, 15317–15331 (2011).
30. Lee, E. B., Skovronsky, D. M., Abtahian, F., Doms, R. W. & Lee, V. M. Secretion and intracellular generation of truncated Aβ in β-site amyloid-β precursor protein-cleaving enzyme expressing human neurons. *J. Biol. Chem.* **278**, 4458–4466 (2003).
31. Richter-Landsberg, C. & Vollgraf, U. Mode of cell injury and death after hydrogen peroxide exposure in cultured oligodendroglia cells. *Exp. Cell Res.* **244**, 218–229 (1998).
32. Duda, J. E. et al. Immunohistochemical and biochemical studies demonstrate a distinct profile of α-Synuclein permutations in Multiple System Atrophy. *J. Neuropathol.* **59**, 830–841 (2000).
33. Luk, K. C. et al. Intracerebral inoculation of pathological α-synuclein initiates a rapidly progressive neurodegenerative α-syncleinopathy in mice. *J. Exp. Med.* **209**, 975–986 (2012).
34. Iba, M. et al. Synthetic tau fibrils mediate transmission of neurofibrillary tangles in a transgenic mouse model of Alzheimer's-like tauopathy. *J. Neurosci.* **33**, 1024–1037 (2013).
35. Abeliovich, A. et al. Mice lacking α-synuclein display functional deficits in the nigrostriatal dopamine system. *Neuron* **25**, 239–252 (2000).

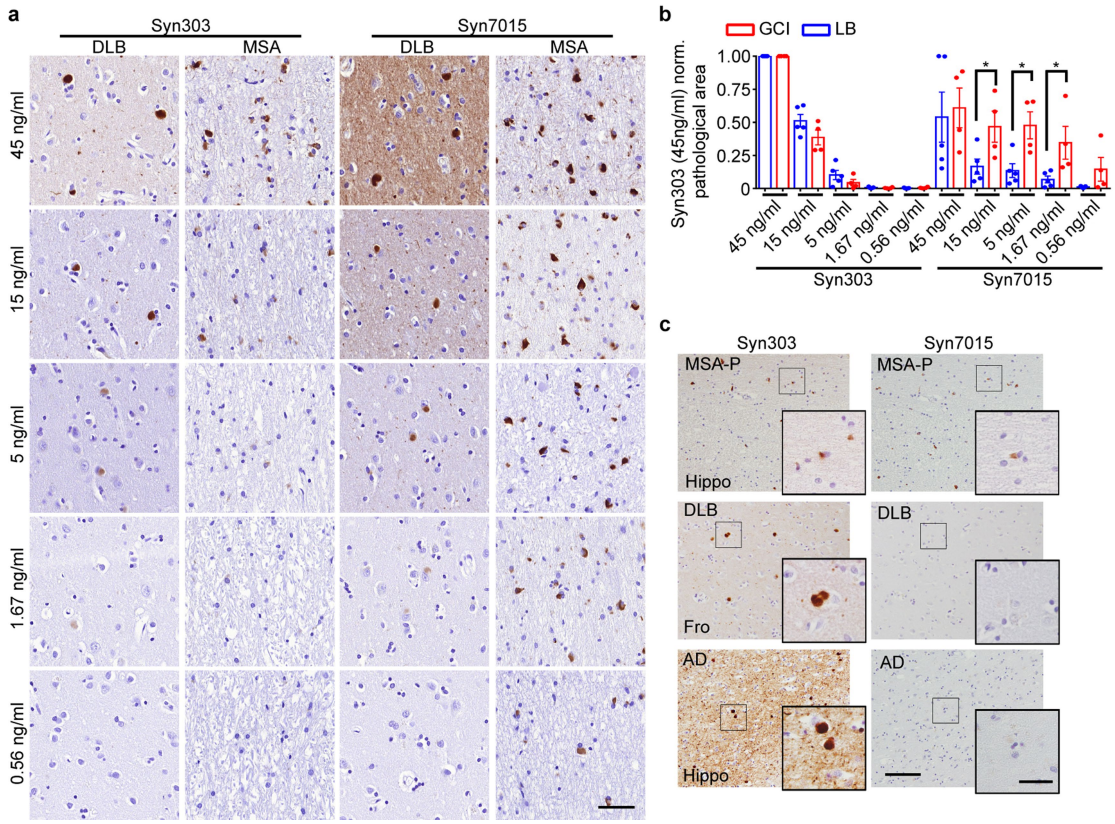
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Antibody	HuA	SNL4	NAC1	Syn204	LB509	Syn211	Syn102	Band
Epitope	1-140	2-11	75-91	87-110	115-122	121-125	131-140	Intensity
Band 1	+	+	+	+	+	+	+	GCI>>LB
Band 2	+	-	+	+	+	+	+/-	GCI<<LB
Band 3	+	+	+	+	+	+	-	GCI<LB
Band 4	+	-	+	+	+	+	-	GCI<LB

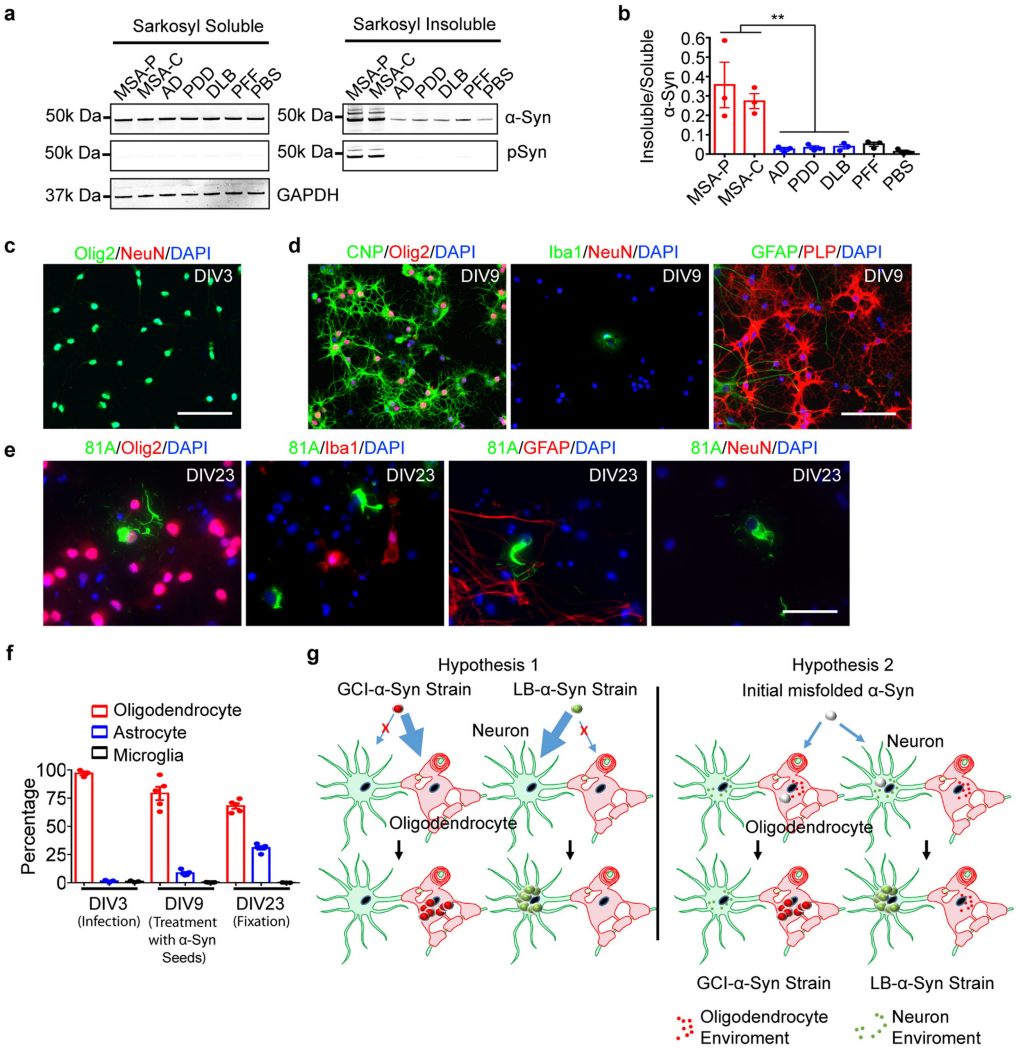
Extended Data Fig. 1 | Biochemical analysis of GCI-α-Syn and LB-α-Syn. a, Schematic for sequential extraction of brains with α-synucleinopathy. Diseased brain samples were sequentially extracted with buffer of increasing extraction strengths (1% Triton X-100 followed by 1% sarkosyl) to remove soluble proteins. **b**, Proteinase K-digested GCI-α-Syn and LB-α-Syn were immunoblotted with a series of antibodies targeting specific domains of α-Syn that spanning the entire molecule. **c**, Summary of the results for experiments described in **b**. **d**, Thermolysin-digested and undigested sarkosyl-insoluble fractions from three cases of Lewy body disease (LB1–LB3) and three cases of MSA (GCI1–GCI3) were resolved on 12% Bis-Tris gel and immunoblotted with an antibody against α-Syn (Syn211). **e**, Sarkosyl-insoluble fractions from a pair of cases (one of Lewy body disease and one of MSA) were incubated with increasing concentrations of thermolysin (with a ratio of thermolysin to total protein that ranged from 1.25×10^{-2} to 5×10^{-2}) and immunoblotted with antibody against α-Syn (Syn211). Undigested fractions were loaded on the same gel. **f**, Trypsin-digested and undigested sarkosyl-insoluble fractions from three cases of Lewy body disease (LB1–LB3) and three cases of MSA (GCI1–GCI3) were resolved on 12% Bis-Tris gel and immunoblotted with an antibody against α-Syn (Syn211). **g**, Sarkosyl-insoluble fractions from a pair of cases (one of Lewy body disease and one of MSA) were incubated with increasing concentrations of trypsin (with the ratio of trypsin versus total protein range from 1.25×10^{-2} to 5×10^{-2}) and immunoblotted with an antibody against α-Syn (Syn211). Undigested fractions were loaded on the same gel. The experiments shown in **b** and **d**–**g** have been repeated three times with similar results. For gel source data, see Supplementary Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



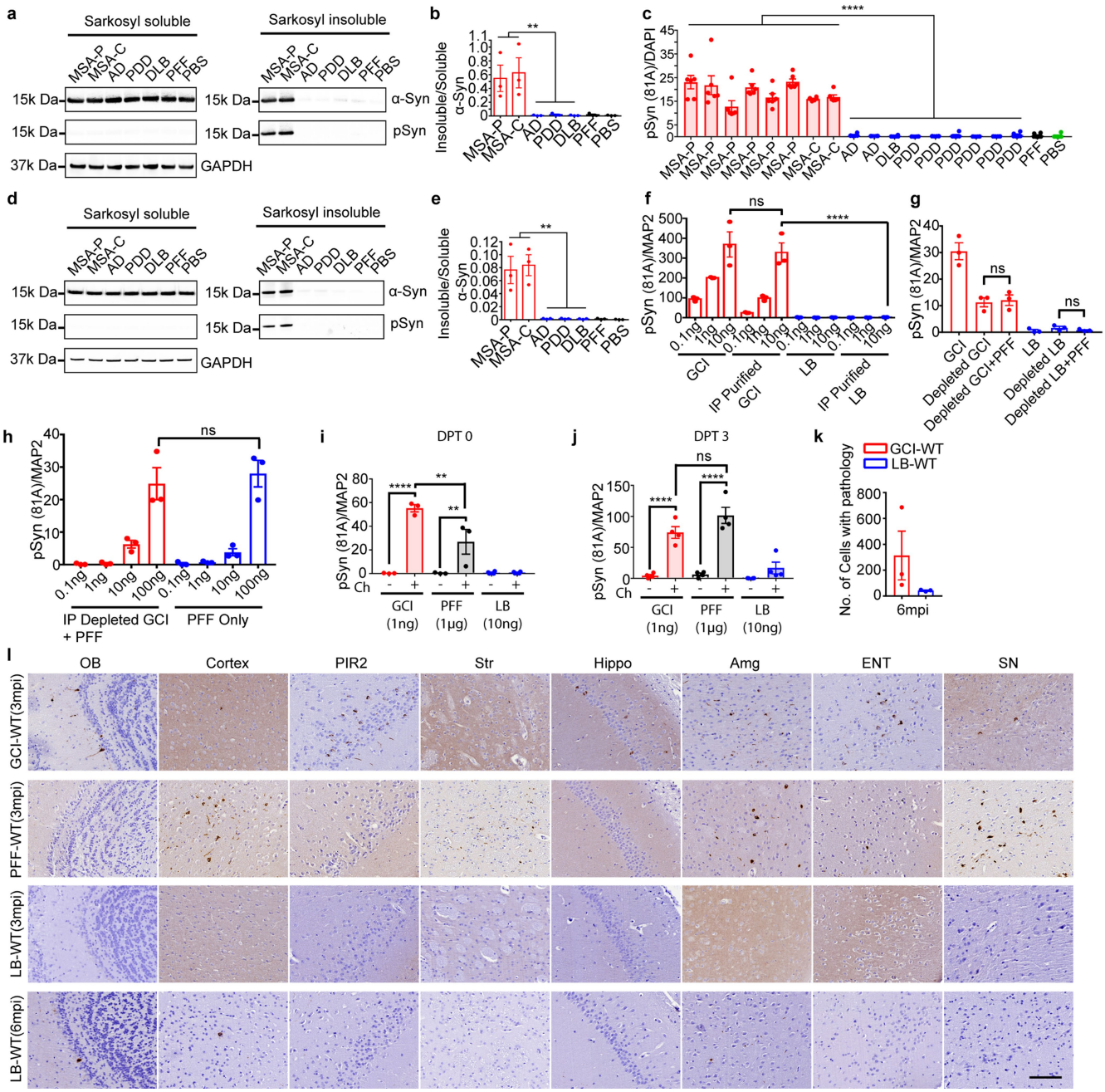
Extended Data Fig. 2 | Syn7015 preferentially recognizes GCIs over Lewy bodies. a, Immunohistochemistry using a series dilution of Syn303 or Syn7015 on serial sections of a DLB brain and a MSA brain. At 45 ng ml^{-1}, Syn7015 recognized both Lewy bodies and GCIs. At lower concentrations—particularly 1.67 ng ml^{-1} and 0.56 ng ml^{-1}—Syn7015 preferentially recognizes GCIs over Lewy bodies (repeated with four cases). **b**, Quantification of the area occupied by pathological α-Syn stained with serial dilutions of Syn7015 or Syn303 on serial sections of two cases of MSA-P, two cases of MSA-C, one case of AD, two cases of PDD and two cases of DLB. The results for each case are normalized to Syn303 staining at 45 ng ml^{-1} (GCI, $n = 4$ cases; LB, $n = 5$ cases). **c**, α-Syn pathology revealed by Syn303 or Syn7015 in adjacent sections from cases of Lewy body disease and MSA (repeated with seven cases). Results shown as mean ± s.e.m. *$P < 0.05$. Scale bar: 50 μm (**a**), 100 μm (**c**), 25 μm (**c** inset). See Supplementary Table 5 for statistical details.Source Data

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 3 | GCI-α-Syn is more potent in inducing α-Syn pathology in primary oligodendrocytes. a, Oligodendrocytes treated with the same amount of GCI-α-Syn, LB-α-Syn or PFF were sequentially extracted with 1% Triton-X100 lysis buffer followed by 1% sarkosyl lysis buffer, which were combined together as the sarkosyl-soluble fraction. The sarkosyl-insoluble pellets were resuspended in Dulbecco's PBS by sonication. Both soluble and insoluble fractions were immunoblotted with an antibody against total or pS129 α-Syn. **b**, Densitometric quantification of insoluble versus soluble α-Syn for experiments described in **a** ($n = 3$ biological replicates). **c**, **d**, Primary oligodendrocyte cultures were immunostained with antibodies against various cell-type specific markers: CNP (oligodendrocytes), olig2 (oligodendrocytes), Iba1 (microglial cells), NeuN (neurons), GFAP (astrocytes), PLP (oligodendrocytes) at day in vitro 3 (DIV 3) (**c**) or DIV 9 (**d**). **e**, Insoluble phosphorylated α-Syn, induced in primary oligodendrocytes overexpressing α-Syn, was co-stained with antibodies against various cell-type specific markers, demonstrating that the cells with α-Syn pathology are oligodendrocytes. **f**, Percentage of different type of cells (oligodendrocytes, microglial cells and astrocytes) in oligodendrocyte culture, at DIV 3 (the time point of virus infection), DIV 9 (the time point for misfolded α-Syn treatment) and DIV 23 (the time point for fixation) ($n = 3$ (DIV3) or 5 (DIV 9, DIV 23) coverslips from three independent experiments). **g**, Working hypotheses regarding the different cell-type distributions of GCI-α-Syn and LB-α-Syn strains in diseased brains. Hypothesis 1 states that the unique properties of GCI-α-Syn and LB-α-Syn strains determine their different cell-type distributions. The GCI-α-Syn strain (represented by red spheres) is more efficient in inducing α-Syn pathology in oligodendrocytes, whereas the LB-α-Syn strain (green spheres) is more efficient in inducing α-Syn pathology in neurons. Hypothesis 2 states that GCI-α-Syn and LB-α-Syn strains do not have cell-type preferences and that they could both be initiated by the same misfolded α-Syn seeds (grey spheres), but that the different intracellular environments of neurons and oligodendrocytes convert these α-Syn seeds to different strains. Results shown as mean ± s.e.m. $^{**}P < 0.01$. Scale bars: 100 μm (**c**, **d**); 50 μm (**e**). The experiments shown in **a** and **c**–**e** have been repeated three times with similar results. See Supplementary Table 5 for statistical details. For gel source data, see Supplementary Fig. 1.Source Data

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

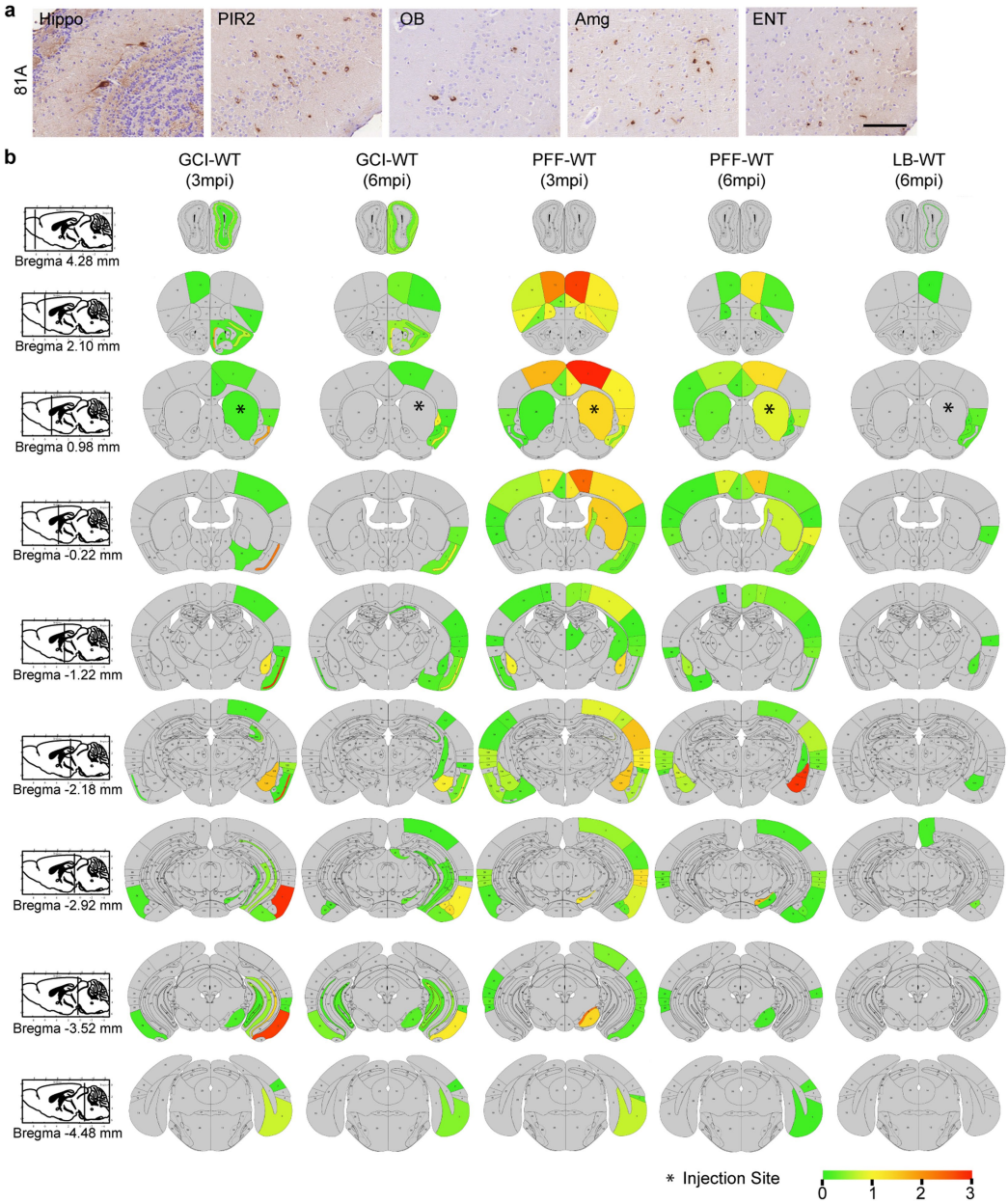


Extended Data Fig. 4 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

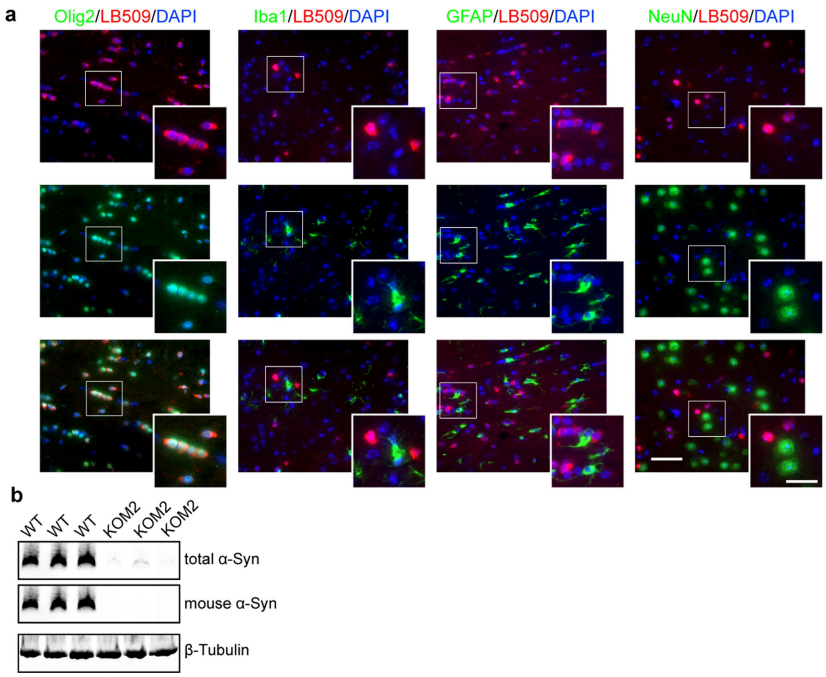
Extended Data Fig. 4 | The seeding properties of GCI-α-Syn and LB-α-Syn do not show cell-type preference. a, Soluble and insoluble fractions from primary neurons treated with the same amount of GCI-α-Syn, LB-α-Syn or PFF were immunoblotted with antibodies against total or pS129 α-Syn. **b**, Densitometric quantification of insoluble versus soluble α-Syn for experiments described in **a** ($n = 3$ biological replicates). **c**, Quantification of phosphorylated α-Syn in QBI-WT-Syn cells induced by an equal amount of GCI-α-Syn (MSA-C, MSA-P), LB-α-Syn (PDD, DLB and AD) or PFFs (GCI, $n = 8$ different preparations; LB, $n = 9$ different preparations). **d**, Soluble and insoluble fractions from QBI-Syn-WT cells treated with the same amount of GCI-α-Syn, LB-α-Syn or PFF were immunoblotted with an antibody against total or pS129 α-Syn. **e**, Densitometric quantification of insoluble versus soluble α-Syn for experiments described in **d** ($n = 3$ biological replicates). **f**, Quantification of insoluble phosphorylated α-Syn in primary neurons induced by various concentrations of GCI-α-Syn and LB-α-Syn before or after immunoprecipitation purification ($n = 3$ independent experiments). **g**, Quantification of insoluble phosphorylated α-Syn in primary neurons incubated with (1) GCI-α-Syn and LB-α-Syn preparations; (2) the same preparations after immunoprecipitation depletion to remove α-Syn; and (3) the depleted preparation to which the same amount of α-Syn PFFs (1 ng) was added ($n = 3$ independent experiments). **h**, PFFs combined with the GCI-α-Syn preparation depleted of α-Syn behave similarly to α-Syn PFFs alone. Quantification of insoluble phosphorylated α-Syn in primary neurons seeded by PFFs alone or PFFs combined with depleted GCI preparation (the amount of pathology induced by immunoprecipitation-depleted GCI preparation alone has been subtracted) ($n = 3$ independent experiments). **i**, **j**, Primary neuron were treated with GCI-α-Syn, LB-α-Syn or PFF and incubated with chloroquine (Ch) at the day of misfolded α-Syn treatment ('DPT0') or three days post-treatment (DPT3). The amount of insoluble phosphorylated α-Syn was quantified three days after chloroquine treatment ($n = 3$ (DPT0-GCI and PFF) or 4 (DPT0-LB, DPT3) independent experiments). **k**, Quantification of the number of cells with α-Syn pathology in wild-type mice inoculated with 50 ng of GCI-α-Syn or LB-α-Syn at six months post-injection. **l**, Representative photomicrographs of α-Syn pathology (stained by Syn506) in multiple brain regions ipsilateral to the injection site in GCI-α-Syn-, PFF- and LB-α-Syn-injected wild-type mice. Cortex, motor cortex; ENT, entorhinal cortex. Results shown as mean ± s.e.m. $*P < 0.05$; $**P < 0.01$; $****P < 0.0001$; ns, not significant. Statistics shown in **c** represent two-tailed, unpaired *t*-test using the mean value of each case. Statistics shown in **f** represent one-way Anova with Tukey's multiple comparison test. Statistics shown in **g**, **h** are two-tailed, unpaired *t*-test adjusted with Bonferroni correction for multiple comparison. Statistics shown in **i**, **j** are two-way ANOVA, with Sidak's multiple comparisons test. The experiments in **a**, **d** and **l** have been repeated three times with similar results. Scale bar: 100 μm. See Supplementary Table 5 for statistical details. For gel source data, see Supplementary Fig. 1.Source Data

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



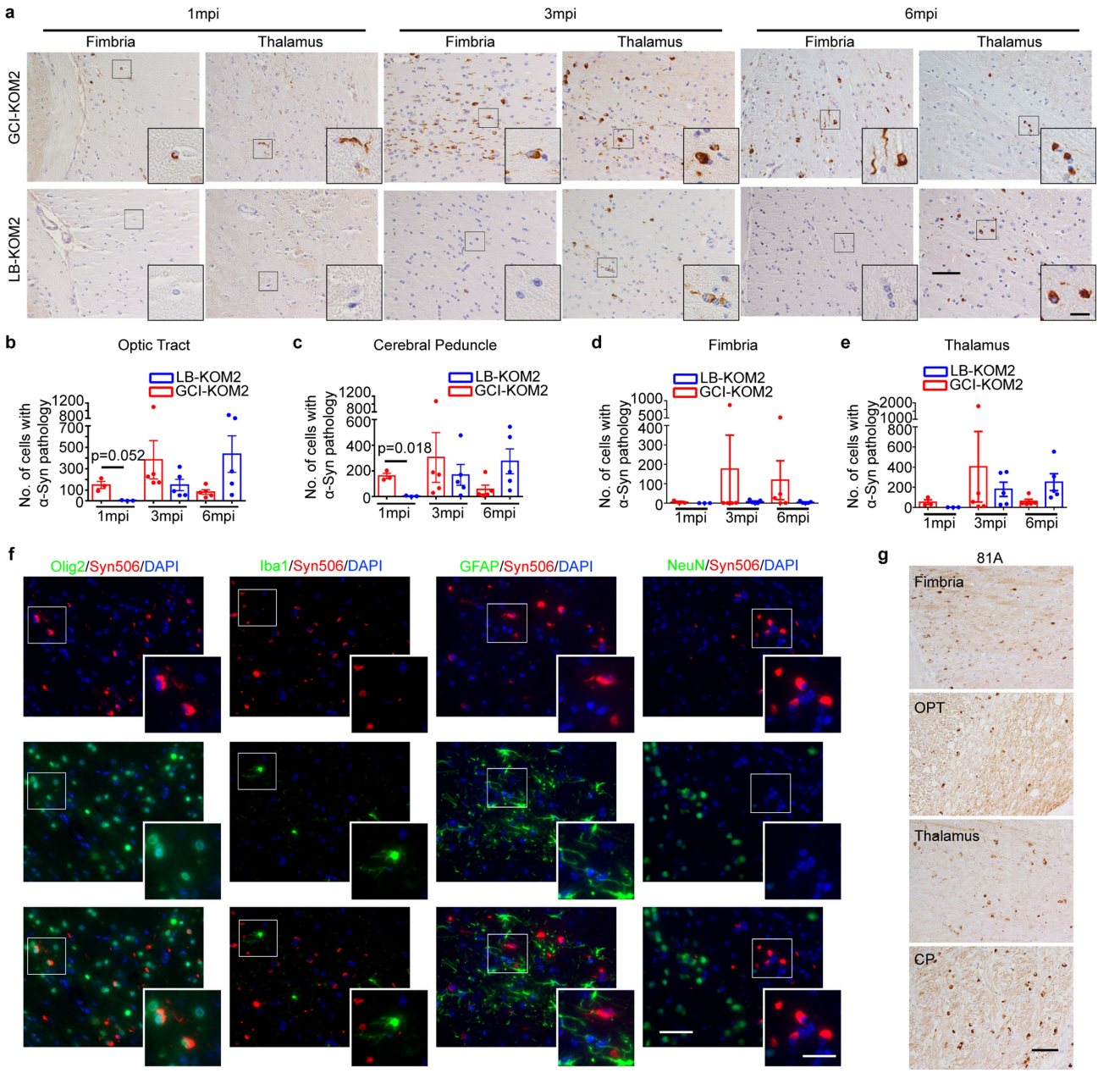
Extended Data Fig. 5 | Distribution of α-Syn pathology in injected wild-type mice. a, Representative photomicrographs of α-Syn pathologies stained by an antibody against pS129 α-Syn (81A) in multiple brain regions in GCI-α-Syn injected wild-type mice (experiment repeated three times). **b**, Heat map for the distribution of α-Syn pathology in wild-type mice injected with GCI-α-Syn, PFF or LB-α-Syn. GCI-α-Syn, LB-α-Syn and α-Syn PFF were unilaterally injected into the dorsal striatum of wild-type mice. The seeded α-Syn pathology was analysed and graded by immunohistochemistry with Syn506. The data were presented as heat maps to semiquantitatively demonstrate the central nervous system (CNS) distribution of α-Syn pathology. Each panel represents a coronal plane (bregma 4.28, 2.10, 0.98, −0.22, −1.22,−2.18, −2.92, −3.52 and −4.48 mm) for each treatment group (GCI-WT, $n = 3$ mice; PFF-WT, $n = 4$ mice; LB-WT, $n = 3$ mice). Scale bar: 100 μm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



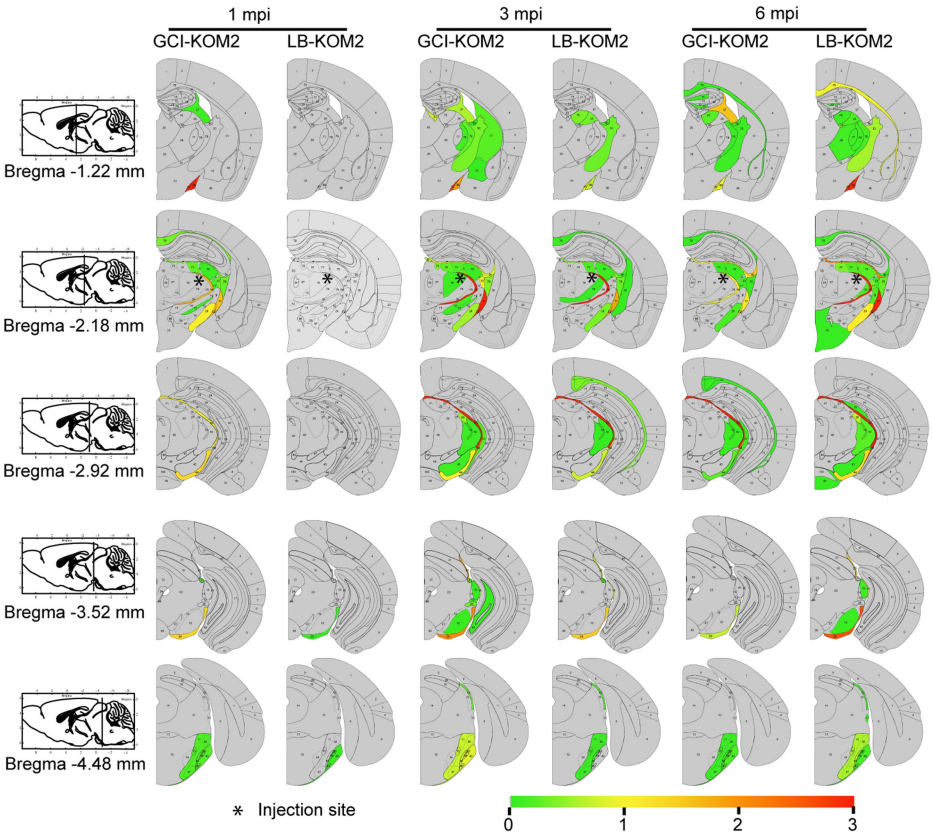
Extended Data Fig. 6 | Characterization of KOM2 mice. a, Brain sections from KOM2 mice were double-labelled with antibodies against α-Syn (LB509) and various cell-type specific markers: Olig2 (oligodendrocytes), Iba1 (microglial cells), GFAP (astrocytes) and NeuN (neurons). In KOM2 mice, α-Syn is expressed only in oligodendrocytes. **b**, Brain lysates of wild-type and KOM2 mice were immunoblotted with an antibody against total α-Syn (Syn 9027), mouse α-Syn (Cell Signalling) and β-tubulin. Scale bars: 50 μm (**a**), 25 μm (**a** inset). The experiments in **a**, **b** have been repeated three times with similar results. For gel source data, see Supplementary Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



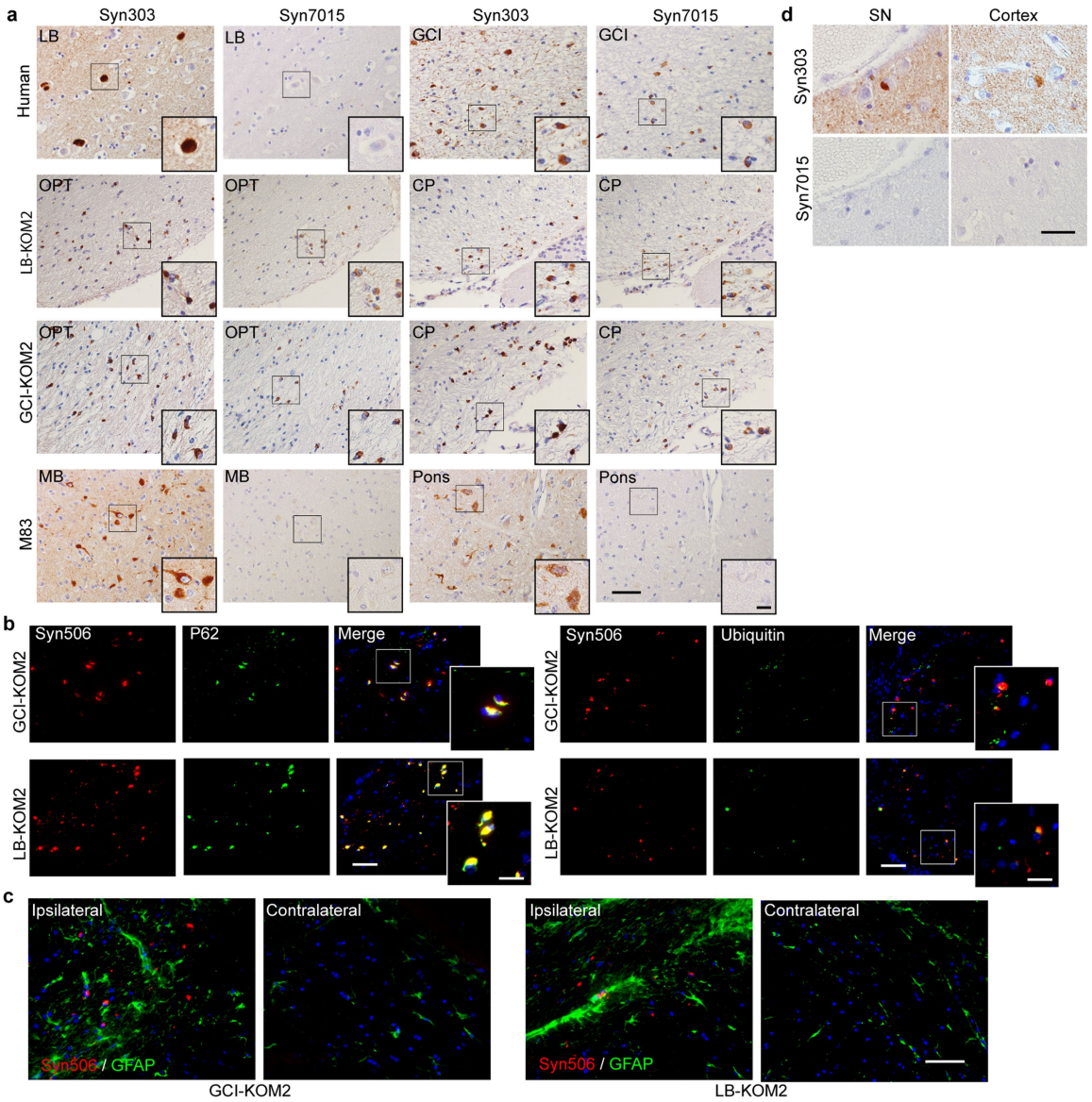
Extended Data Fig. 7 | Induction of oligodendroglial α-Syn pathology in KOM2 mice. a, Syn506-positive α-Syn aggregates seeded by equal amounts of GCI-α-Syn or LB-α-Syn (18.75 ng) in KOM2 mice at one, three and six month post-injection in fimbria and thalamus. **b**–**e**, Quantification of the number of oligodendrocytes with α-Syn pathology in the optic tract (**b**), cerebral peduncle (**c**), fimbria (**d**), and thalamus (**e**) at different time points (one month post-injection, $n = 3$ mice; three and six months post-injection, $n = 5$ mice). **f**, Brain sections from GCI-α-Syn injected KOM2 mice were double-labelled with antibodies against misfolded α-Syn (Syn506) and various cell-type specific markers. The induced α-Syn pathologies are located in oligodendrocytes in KOM2 mice. **g**, GCI-α-Syn-injected KOM2 mouse brain sections were stained with an antibody against phosphorylated α-Syn (81A). Results shown as mean ± s.e.m. Scale bars: 50 μm (**a**, **f** and **g**), 12.5 μm (**a** insets) and 25 μm (**f** insets). The experiments in **a**, **f** and **g** have been repeated three times with similar results. Statistics shown in **b**, **c** are two tailed unpaired *t*-test adjusted with Bonferroni correction. See Supplementary Table 5 for statistical details.Source Data

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



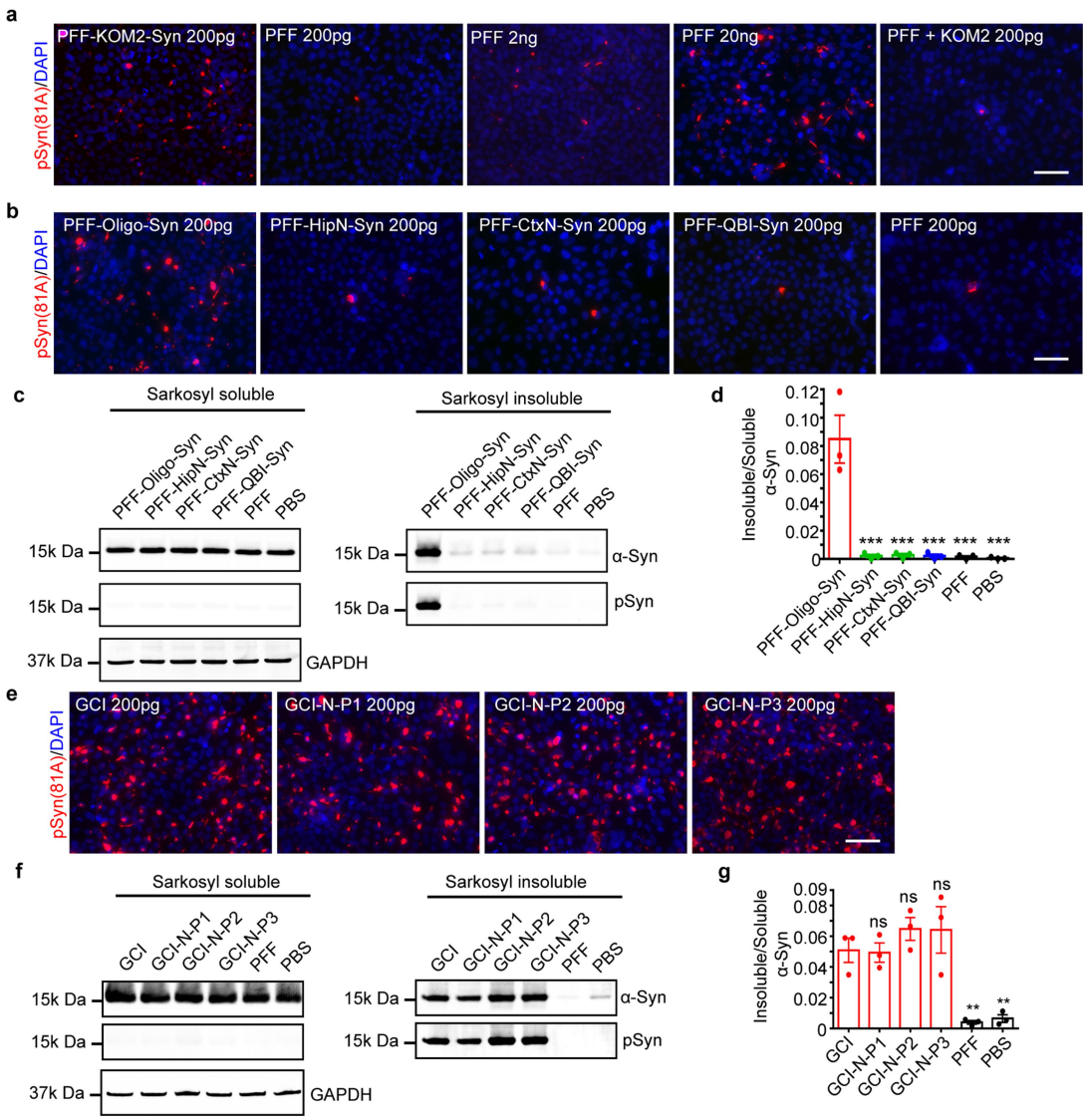
Extended Data Fig. 8 | Distribution of α-Syn pathology in injected KOM2 mice. Heat maps to semiquantitatively demonstrate the CNS distribution of α-Syn pathology in KOM2 mice unilaterally injected with GCI-α-Syn or LB-α-Syn into the thalamus. α-Syn pathologies were analysed and graded by immunohistochemistry with Syn506 at one month ($n = 3$), three months ($n = 5$) and six month ($n = 5$) post injection. Each panel represents a coronal plane (bregma −1.22, −2.18, −2.92, −3.52, −4.48 mm) for each treatment group. Because there is no α-Syn pathology in the contralateral side, only the ipsilateral side is shown.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | Oligodendrocyte environment generates the GCI-α-Syn strain. a, Immunohistochemistry of adjacent sections from human or mouse brains with Syn303 and Syn7015. First row, adjacent brain sections of cases of Lewy body disease and MSA used for the extraction of LB-α-Syn and GCI-α-Syn for injection. Second row, adjacent brain sections of KOM2 mice injected with LB-α-Syn prepared from the brain tissue shown in the first row. OPT and CP are shown. Whereas the LB-α-Syn used for the injections is Syn7015-negative, the oligodendrocyte pathology that is induced is Syn7015-positive. Third row, adjacent brain sections of KOM2 mice injected with GCI-α-Syn prepared from the brain sample shown in the first row. OPT and CP are shown. Fourth row, adjacent brain sections of M83 mice with α-Syn pathology. Midbrain (MB) and pons are shown. **b**, Brain sections from KOM2 mice injected with GCI-α-Syn or LB-α-Syn in the thalamus were double-labelled with Syn506 and antibodies against P62 (left panel) or ubiquitin (right panel). **c**, Brain sections from GCI-α-Syn- or LB-α-Syn-injected KOM2 mice were double-labelled with Syn506 and GFAP. Both ipsilateral and contralateral optic tracts are shown. **d**, Adjacent sections of substantia nigra and cortex from two different cases of MSA were stained with Syn7015 and Syn303. Scale bars: 50 μm (**a**–**c**), 12.5 μm (**a** insets), 20 μm (**b** inset) and 30 μm (**d**). The experiments in **a**–**d** have been repeated three times with similar results.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | α-Syn pathology induced by passaged PFF and GCI. a, Insoluble phosphorylated α-Syn in QBI-WT-Syn cells seeded by PFFs, PFFs that have been passaged in KOM2 mice (PFF-KOM2-Syn) or PFFs that were combined with the sarkosyl-insoluble fraction prepared from uninjected KOM2 mice (PFF + KOM2). **b**, Insoluble phosphorylated α-Syn in QBI-WT-Syn cells induced by an equal amount (200 pg) of PFF-oligo-Syn, PFF-HipN-Syn, PFF-CtxN-Syn, PFF-QBI-Syn and PFFs. **c**, Soluble and insoluble fractions from QBI-Syn-WT cells treated with the same amount of PFF-oligo-Syn, PFF-HipN-Syn, PFF-CtxN-Syn, PFF-QBI-Syn and PFF were immunoblotted with antibodies against total or pS129 α-Syn. **d**, Densitometric quantification of insoluble versus soluble α-Syn for experiments described in **c** ($n = 3$ biological replicates). **e**, Insoluble phosphorylated α-Syn in QBI-WT-Syn cells induced by GCI-α-Syn and GCI-α-Syn that has been passaged in primary neurons for multiple times (that is, GCI-N-P1, GCI-N-P2, GCI-N-P3). **f**, Soluble and insoluble fractions from QBI-Syn-WT cells treated with the same amount of GCI, GCI-N-P1, GCI-N-P2, GCI-N-P3 and PFF were immunoblotted with antibodies against total α-Syn or pS129 α-Syn. **g**, Densitometric quantification of insoluble versus soluble α-Syn for experiments described in **f** ($n = 3$ biological replicates). Statistics shown in **d**, **g** are one-way ANOVA followed by Dunnett's post hoc test comparing each group with PFF-oligo-Syn in **d** or GCI in **g**. Results shown as mean ± s.e.m. $^{**}P < 0.01$; $^{***}P < 0.01$; ns, not significant. Scale bars: 50 μm (**a**, **b**, **e**). The experiments in **a**–**c** and **e**–**f** have been repeated three times with similar results. See Supplementary Table 5 for statistical details. For gel source data, see Supplementary Fig. 1.Source Data

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0118-y

RSPO2 inhibition of RNF43 and ZNRF3 governs limb development independently of LGR4/5/6

Emmanuelle Szenker-Ravi[1,17], Umut Altunoglu[2,17], Marc Leushacke[1,17], Célia Bosso-Lefèvre[1,3], Muznah Khatoo[1], Hong Thi Tran[4], Thomas Naert[4], Rivka Noelanders[4], Amin Hajamohideen[1], Claire Beneteau[5], Sergio B. de Sousa[6,7], Birsen Karaman[2], Xenia Latypova[5], Seher Başaran[2], Esra Börklü Yücel[8], Thong Teck Tan[1], Lena Vlaeminck[4,16], Shalini S. Nayak[9], Anju Shukla[9], Katta Mohan Girisha[9], Cédric Le Caignec[5,10], Natalia Soshnikova[11], Zehra Oya Uyguner[2], Kris Vleminckx[4]*, Nick Barker[1,12,13]*, Hülya Kayserili[2,8]* & Bruno Reversade[1,3,8,14,15]*

The four R-spondin secreted ligands (RSPO1–RSPO4) act via their cognate LGR4, LGR5 and LGR6 receptors to amplify WNT signalling[1–3]. Here we report an allelic series of recessive *RSPO2* mutations in humans that cause tetra-amelia syndrome, which is characterized by lung aplasia and a total absence of the four limbs. Functional studies revealed impaired binding to the LGR4/5/6 receptors and the RNF43 and ZNRF3 transmembrane ligases, and reduced WNT potentiation, which correlated with allele severity. Unexpectedly, however, the triple and ubiquitous knockout of *Lgr4*, *Lgr5* and *Lgr6* in mice did not recapitulate the known *Rspo2* or *Rspo3* loss-of-function phenotypes. Moreover, endogenous depletion or addition of exogenous RSPO2 or RSPO3 in triple-knockout *Lgr4/5/6* cells could still affect WNT responsiveness. Instead, we found that the concurrent deletion of *rnf43* and *znrf3* in *Xenopus* embryos was sufficient to trigger the outgrowth of supernumerary limbs. Our results establish that RSPO2, without the LGR4/5/6 receptors, serves as a direct antagonistic ligand to RNF43 and ZNRF3, which together constitute a master switch that governs limb specification. These findings have direct implications for regenerative medicine and WNT-associated cancers.

Limb development is governed by a three-dimensional signalling system that defines proximodistal, anteroposterior and dorsoventral axes[4]. Tetra-amelia with lung hypo/aplasia syndrome (TETAMS; MIM database entry 273395) is an extreme condition, in which fetuses lack all four limbs. TETAMS without lung hypoplasia has been linked to a *WNT3* nonsense mutation in humans[5]. The four R-spondin ligands (RSPO1–RSPO4) act as enhancers of WNT signalling[6–9]. They bind to their cognate receptors LGR4, LGR5 and LGR6 via their Furin-like 2 domain, and to the E3 ubiquitin ligases RNF43 and ZNRF3 via their Furin-like 1 domain[10]. This tripartite interaction prevents WNT receptor degradation mediated by RNF43 or ZNRF3. *Rspo2* mutation in mice leads to limb truncations reminiscent of tetra-amelia[11–14], but a role for its receptors has not been substantiated, to our knowledge, during limb morphogenesis.

Here, we describe five families with eleven affected individuals that display severe developmental limb defects. In family 1, four affected fetuses presented with radial ray deficiency with humeral involvement, absence of tibiae with or without femoral deficiency, and absence of digits on the preaxial side (Fig. 1a, b, Extended Data Fig. 1a and Extended Data Table 1). We propose to name this severe dysostosis humero-femoral hypoplasia with radio-tibial ray deficiency (HFH-RTRD). Exome sequencing identified a homozygous p.Arg69Cys mutation in *RSPO2* (Extended Data Fig. 1b) that affects a residue conserved in all R-spondin paralogues and homologues (Fig. 1c and Extended Data Fig. 2a). The analogous mutation p.Arg64Cys in *RSPO4* was shown to cause congenital anonchia[15]. The seven affected fetuses from families 2 to 5 presented with complete absence of four limbs, lung hypo/aplasia, cleft lip-palate, and labioscrotal fold aplasia, all characteristic of TETAMS (Fig. 1a, b, Extended Data Fig. 1a and Extended Data Table 1). A p.Gln70* nonsense mutation in *RSPO2* was identified in family 2[16] (Fig. 1a, c). In family 3[17], array comparative genomic hybridization (array-CGH) analysis identified a biallelic deletion of 154 kilobases (kb) spanning intron 5 and exon 6 of *RSPO2* (Fig. 1a, c and Extended Data Fig. 1c). In family 4, exome sequencing revealed a recessive p.Glu137* nonsense mutation in *RSPO2* (Fig. 1a, c). In family 5 with three consecutive TETAMS fetuses, a homozygous *RSPO2* frameshift p.Gly42Val fs*49 mutation was identified (Fig. 1a–c). These results establish a new aetiology for tetra-amelia and demonstrate the crucial involvement of RSPO2 in craniofacial, limb and lung development in humans.

We selected the p.Arg69Cys (R69C) and p.Gln70* (Q70X) mutations that are responsible for HFH-RTRD and TETAMS, respectively, to assess whether mutant RSPO2 retained binding to its cognate receptors. By co-immunoprecipitation analysis, only wild-type RSPO2, but not the RSPO2(F105A/F109A) mutant that specifically abrogates binding to LGRs[18], nor the RSPO2(R69C) or RSPO2(Q70X) mutants, was able to be pulled down by LGR5 (Fig. 1d and Extended Data Fig. 2b, c). Although the RSPO2(F105A/F109A) mutant could be readily co-immunoprecipitated by RNF43 or ZNRF3, neither RSPO2(R69C) nor RSPO2(Q70X) could interact with RNF43 or ZNRF3 (Fig. 1e and Extended Data Fig. 2d). Similarly, only wild-type RSPO2, but not the RSPO2(R69C) or RSPO2(Q70X) mutants, could be retained on the surface of HEK293T cells overexpressing LGR5 or RNF43 (Extended Data Fig. 2e). Wild-type RSPO2, and to a lesser extent RSPO2(R69C) but not RSPO2(Q70X), could enhance WNT3A-mediated activation of SUPERTOPFLASH (STF) luciferase (Fig. 1f and Extended Data Fig. 2f). These in vitro results indicate that the R69C and Q70X mutations diminish the ability of RSPO2 to bind to LGRs, RNF43 or ZNRF3, and to amplify β-catenin-dependent WNT signalling. These signalling defects correlate with the severity of the fetuses' phenotypes—the nonsense Q70X mutation (responsible for TETAMS) behaving as a null mutation, and the R69C mutation (responsible for HFH-RTRD) as a hypomorphic allele.

In mice, *Rspo2* is expressed in the apical ectodermal ridge of the growing limb and in the lung mesenchyme[11]. Accordingly, *Rspo2*

[1]Institute of Medical Biology, A*STAR, Singapore, Singapore. [2]Medical Genetics Department, Istanbul Medical Faculty, Istanbul University, Istanbul, Turkey. [3]Department of Paediatrics, National University of Singapore, Singapore, Singapore. [4]Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium. [5]CHU Nantes, Service de Génétique Médicale, Nantes, France. [6]Medical Genetics Unit, Hospital Pediátrico, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal. [7]University Clinic of Genetics, Faculty of Medicine, University of Coimbra, Coimbra, Portugal. [8]Medical Genetics Department, Koç University School of Medicine (KUSOM), Istanbul, Turkey. [9]Department of Medical Genetics, Kasturba Medical College, Manipal University, Manipal, India. [10]INSERM, UMR1238, Bone Sarcoma and Remodeling of Calcified Tissue, Université Bretagne Loire, Nantes, France. [11]Institute of Molecular Biology (IMB) gGmbH, Mainz, Germany. [12]Cancer Research Institute, Kanazawa University, Kakuma-machi, Kanazawa, Japan. [13]Centre for Regenerative Medicine, The University of Edinburgh, Edinburgh, UK. [14]Institute of Molecular and Cellular Biology, A*STAR, Singapore, Singapore. [15]Reproductive Biology Laboratory, Academic Medical Center (AMC), Amsterdam-Zuidoost, The Netherlands. [16]Present address: Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [17]These authors contributed equally: Emmanuelle Szenker-Ravi, Umut Altunoglu, Marc Leushacke.
*e-mail: kris.vleminckx@irc.ugent.be; nicholas.barker@imb.a-star.edu.sg; hkayserili@kuh.ku.edu.tr; bruno@reversade.com

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

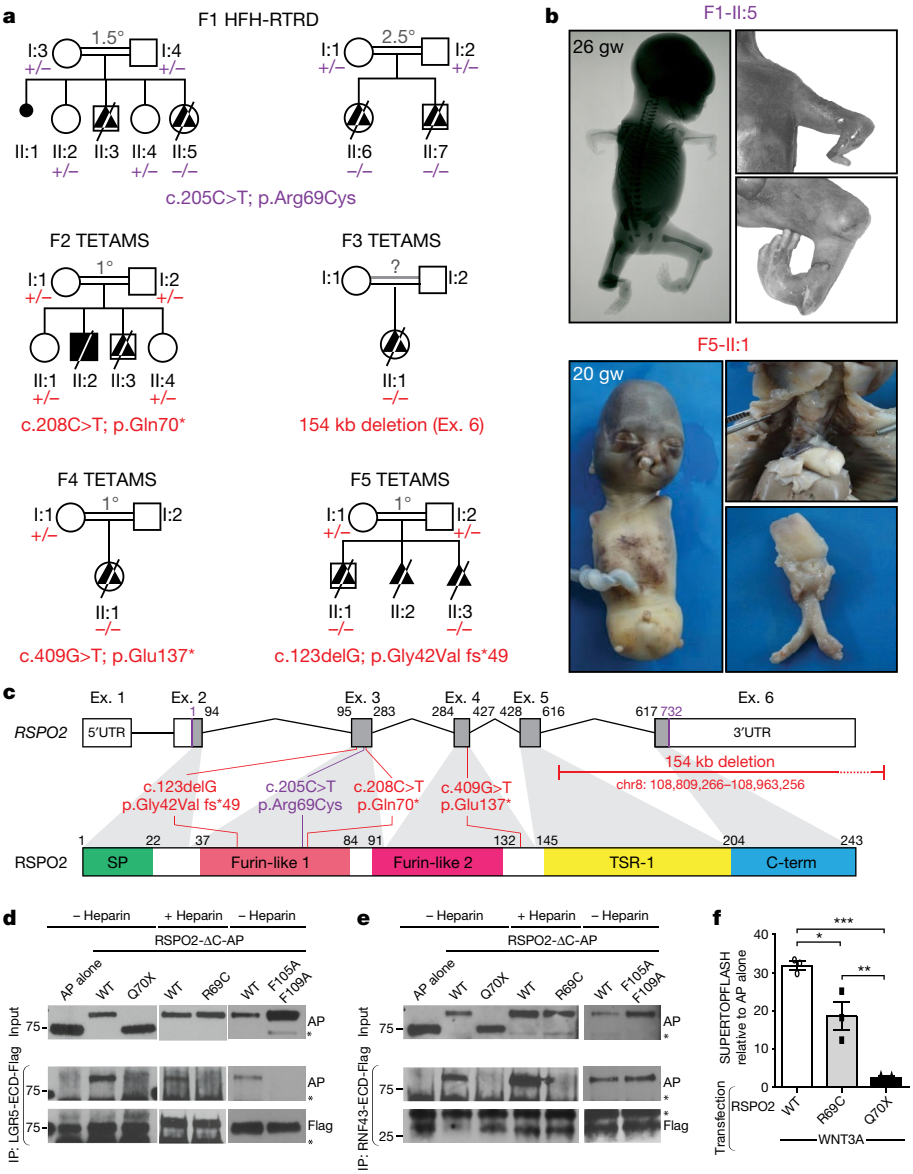


Fig. 1 | Identification of *RSPO2* mutations in fetuses presenting with severe limb defects. a, Pedigrees of family 1 (F1) with HFH-RTRD and families 2–5 (F2–F5) with TETAMS, and *RSPO2* germline mutation status for available family members. **b**, Pictures and radiographs illustrating limb defects in a fetus with HFH-RTRD, and complete absence of limbs and lungs in a fetus with TETAMS. gw, gestational weeks. **c**, *RSPO2* genomic (top) and protein (bottom) structures with identified mutations. Ex., exon. UTR, untranslated region. **d**, Co-immunoprecipitation (IP) of alkaline phosphatase (AP)-tagged RSPO2 lacking the C-terminal domain (ΔC; RSPO2-ΔC-AP) with Flag-tagged extracellular domains (ECD) of LGR5 (LGR5-ECD-Flag). WT, wild-type. **e**, Co-immunoprecipitation of RSPO2-ΔC-AP with RNF43-ECD-Flag. Asterisks indicate non-specific bands. Experiments in **d** and **e** were repeated three times. **f**, SUPERTOPFLASH assay in HEK293T-STF cells transfected with WNT3A and the indicated RSPO2 constructs. $n = 3$ biological replicates. Data are mean ± s.e.m. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, one-way analysis of variance (ANOVA) with Bonferroni's correction. For gel source data, see Supplementary Fig. 1

homozygous mutant mice show lung hypoplasia and limb truncations[11–14]. Consistent with a reduction in canonical WNT signalling, these *Rspo2* phenotypes could be recapitulated in embryos born to gestating mice fed with a pan-WNT inhibitor[19] (Fig. 2a–d). Because individual *Lgr4*, *Lgr5* or *Lgr6* mutant mice[20–22] and *LGR4* or *LGR6* human knockout individuals[23,24] do not present any limb or lung phenotypes, we surmised that functional redundancy might exist between these three receptors. We therefore set to recapitulate TETAMS in mice by genetically deleting *Lgr4/5/6* in all embryonic tissues (Extended Data Fig. 4a). $Lgr4^{+/-}Lgr5^{+/-}Lgr6^{-/-}$ animals were inbred, yielding a 1 in 16 chance of obtaining triple-knockout *Lgr4/5/6* offspring (Extended Data Fig. 3a). Five triple-knockout *Lgr4/5/6* embryos were obtained at embryonic day (E) 14.5 or E18.5. Although they displayed the expected *Lgr4* and *Lgr5* phenotypes (Extended Data Fig. 4b–e)[20,21,25], none exhibited phenotypes reminiscent of tetra-amelia with lung agenesis (Fig. 2e–g and Extended Data Fig. 3b, c). This suggests that the LGR4, LGR5 and LGR6 receptors are not functionally redundant and do not mediate RSPO2 signalling for limb and lung morphogenesis. Other developmental phenotypes such as cleft palate[26] and ankyloglossia[21] were common between *Rspo2* and triple-knockout *Lgr4/5/6* embryos (Extended Data Figs. 3d, 4e, Extended Data Table 1). Notably, *Rspo3* knockout mouse embryos die at E10.5 owing to defective vascularization[27], a phenotype not seen in *Lgr4/5/6* triple-knockout embryos (Extended Data Fig. 3e). Thus, RSPO3-mediated vascularization is also LGR4/5/6-independent. These in vivo genetic findings suggest that RSPO2 and RSPO3 may engage other receptors for limb, lung and vascular development.

The expression of *Lgr4/5/6*, *Rnf43* and *Znrf3* could be examined in the recovered mutant embryos at E14.5. As expected, they had no *Lgr4* or *Lgr5* expression (Extended Data Fig. 5a–e). Some residual

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

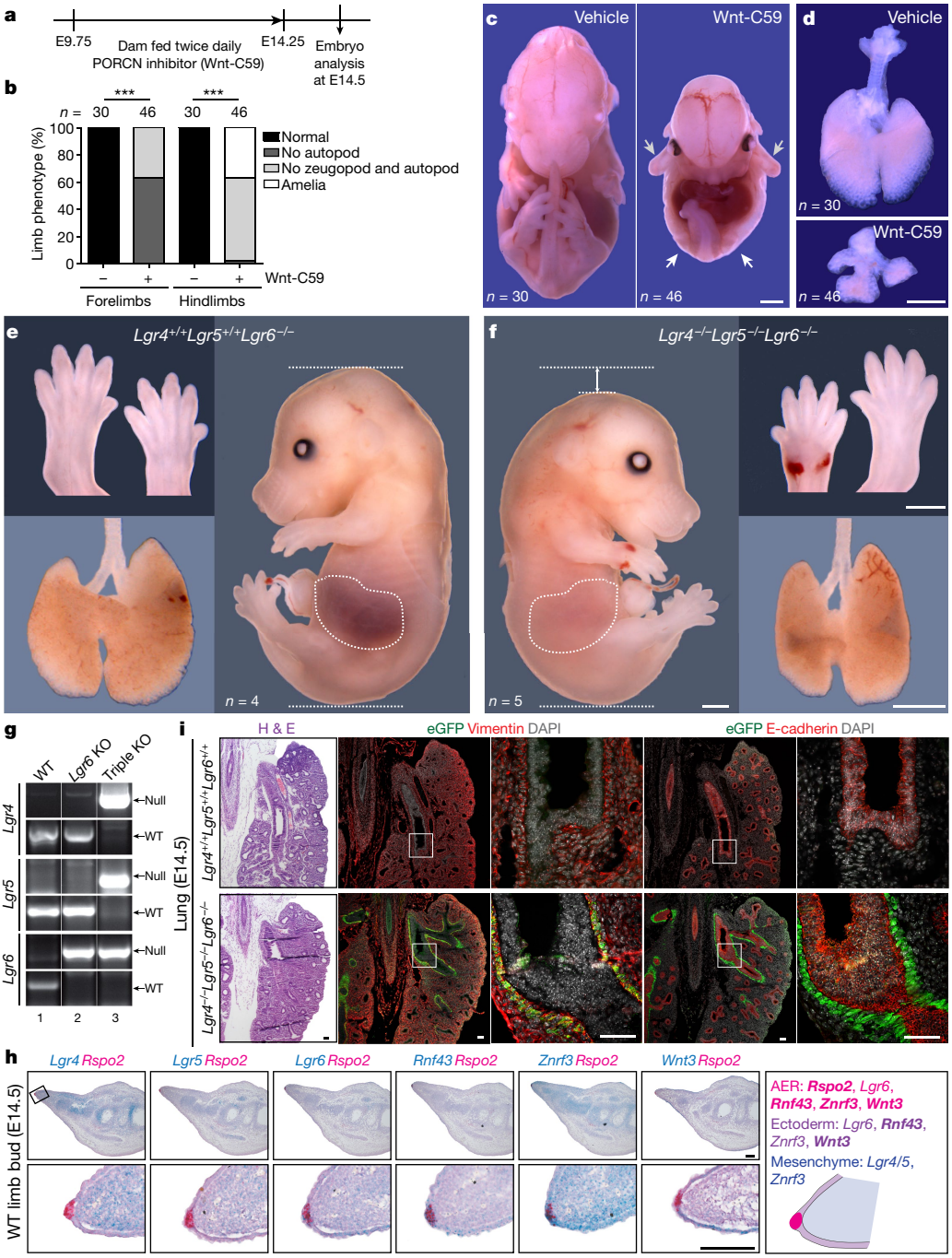


Fig. 2 | Mouse *Lgr4/5/6* triple-knockout embryos do not recapitulate the *Rspo2* and *Rspo3* phenotypes. a–d, PORCN inhibition using Wnt-C59 during embryogenesis leads to limb and lung defects. **a**, Experimental scheme. **b**, Scoring of limb phenotypes in vehicle-treated ($n = 30$) and Wnt-C59 treated ($n = 46$) embryos. ***$P < 0.001$, two-sided Fisher's exact test. **c**, Representative images of treated embryos. Grey (no zeugopod and autopod), and white (amelia) arrows denote limb defects. Scale bars, 1 mm. **d**, Representative images of lungs from treated embryos. Scale bar, 1 mm. **e–g**, Triple-knockout *Lgr4/5/6* does not lead to limb or lung defects. Representative photos of a *Lgr6* knockout ($n = 4$) (**e**) and a triple-null *Lgr4/5/6* ($n = 5$) (**f**) embryo at E14.5, with intact limbs and lungs. Dotted lines indicate the size difference and expected liver position. Scale bars, 1 mm. **g**, PCR-based genotyping. KO, knockout. **h**, Duplex RNAscope images for the indicated transcripts (blue) and *Rspo2* (pink) in transverse sections of wild-type forelimbs. AER, apical ectodermal ridge. Strongly expressed genes are denoted in bold (summary on the right). Scale bars, 0.1 mm. **i**, Haematoxylin and eosin (H&E) and antibody staining in coronal sections of wild-type (top) and triple-knockout (bottom) *Lgr4/5/6* lungs. Scale bars, 50 μm. Experiments in **h** and **i** were repeated three times

Lgr6 expression was observed, which may be originating from an alternative downstream methionine that would delete the LGR6 signal peptide[28]. During limb development, *Lgr4* and *Lgr5* were not expressed in the overlying ectoderm of the limb bud, whereas ectodermal *Lgr6* and *Wnt3* co-localized with *Rspo2* in the apical ectodermal ridge. The expression of *Znrf3* was ubiquitous, whereas *Rnf43* was restricted to the ectoderm (Fig. 2h). In developing lungs, robust *Lgr6* expression was detected in the smooth muscle cells (SMCs), whereas *Lgr4* and *Lgr5* were expressed at low levels in both the epithelium and mesenchyme lineages. *Znrf3* was ubiquitous and overlapped with *Rspo2* in the mesenchyme, whereas *Rnf43* expression was restricted to the lung epithelium (Extended Data Fig. 5f). Using the enhanced green fluorescent protein (eGFP) reporter of the *Lgr4/5/6* knock-in alleles (Extended Data Fig. 4a), we confirmed eGFP expression in a single

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

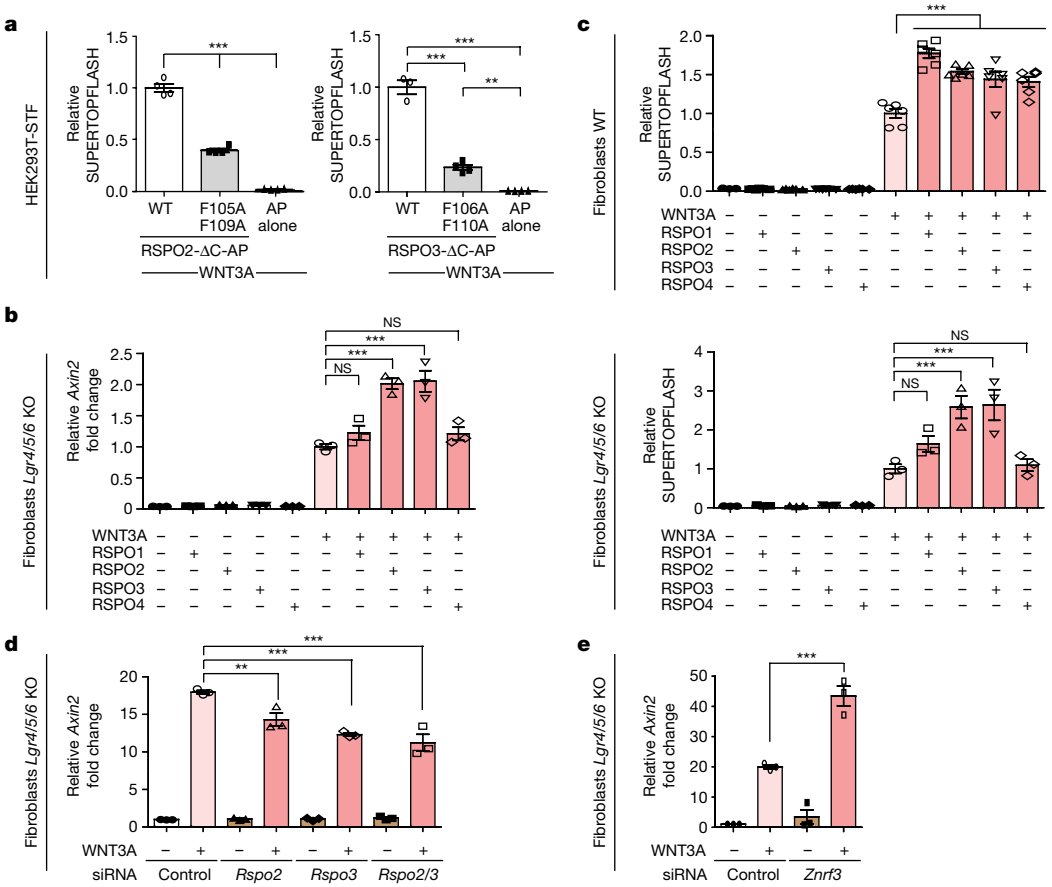


Fig. 3 | Exogenous and endogenous RSPO2/3 signal in *Lgr4/5/6* triple-knockout mouse embryonic fibroblasts. a, SUPERTOPFLASH assay in HEK293T-STF cells transfected with WNT3A and the indicated RSPO2 or RSPO3 constructs. $n = 4$ biological replicates. **b**, qPCR analysis for *Axin2* in *Lgr4/5/6* triple-knockout SV40-immortalized mouse fibroblasts treated with WNT3A and/or RSPO1–RSPO4. $n = 3$ biological replicates. **c**, SUPERTOPFLASH assay in wild-type (top, $n = 6$) and *Lgr4/5/6* triple-knockout (bottom, $n = 3$ biological replicates) SV40-immortalized mouse fibroblasts treated with WNT3A and/or RSPO1–RSPO4. **d**, **e**, qPCR analysis for *Axin2* in *Lgr4/5/6* triple-knockout SV40-immortalized mouse fibroblasts transfected with the indicated siRNAs, and treated with or without WNT3A. $n = 3$ biological replicates. Data are mean ± s.e.m. NS, not significant. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, one-way ANOVA with Bonferroni's correction

layer of vimentin-positive cells adjacent to the E-cadherin-positive lung epithelium (Fig. 2i). In summary, these results demonstrate consistent co-expression of *Rspo2* with *Znrf3* at E14.5, whereas only partial overlap with *Lgr4/5/6* was seen.

We noticed that mutant RSPO2(F105A/F109A) and RSPO3(F106A/F110A), which cannot bind LGRs, are still able to enhance WNT signalling in HEK293T-STF cells (Fig. 3a). Neural progenitor cells (NPCs), induced pluripotent stem (iPS) cells and SV40-immortalized dermal fibroblasts were derived from E14.5 wild-type and mutant embryos to test the activity of exogenous and endogenous R-spondin ligands in triple *Lgr4/5/6*-null cells (Extended Data Fig. 6a–c). Recombinant RSPO2 and RSPO3, but not RSPO1 and RSPO4, could still amplify WNT3A-mediated signalling in *Lgr4/5/6* triple-knockout immortalized STF-fibroblasts (Fig. 3b, c). Most importantly, short interfering RNA (siRNA)-mediated depletion of endogenous *Rspo2* or *Rspo3* was sufficient to significantly decrease expression of the WNT direct target gene *Axin2* in WNT3A-treated *Lgr4/5/6* triple-knockout fibroblasts. This may be explained by RSPO2 and RSPO3 inhibition of ZNRF3, because siRNA-depletion of endogenous *Znrf3* resulted in increased endogenous *Axin2* expression (Fig. 3d, e and Extended Data Fig. 6d). These in vitro data support our in vivo results, and confirm that cells that lack LGR4/5/6 are still sensitive to RSPO2/3-mediated WNT signalling enhancement. Similar observations were made in human haploid cells mutant for LGR4/5/6[29].

To further validate the causal link between RSPO2 deficiency and amelia, we unilaterally injected *rspo2* guide RNA (gRNA) with Cas9 protein into *Xenopus tropicalis* embryos at the two-cell stage (Fig. 4a). Targeted next-generation sequencing and BATCH-GE analysis demonstrated very high in vivo efficiencies for *rspo2* deletions (Supplementary Table 1), which caused marked unilateral forelimb and hindlimb amelia (Fig. 4b, d). Because we showed in mice that LGR4, LGR5 and LGR6 are not involved in limb development, we examined RNF43 and ZNRF3, which may serve as alternative cell-surface RSPO2 receptors. The use of *Xenopus* allows us to bypass a possible mammalian-specific RNF43 and ZNRF3 requirement for placental vascularization. Both ligases were uniformly expressed in developing limb buds (Fig. 4c). Two TALEN pairs for each gene were selected for their very high cutting efficiency (Supplementary Table 1). Although limb defects were rare, or absent, within single *rnf43* or *znrf3* mutants, unilateral ectopic limbs were very prominent in *znrf3/rnf43* double-mutant frogs (Fig. 4d–h). Alizarin red and alcian blue staining revealed a diverse spectrum of limb phenotypes ranging from diplopodia to complete polymelia, with bifurcations arising at distinct locations across the stylopod, zeugopod or autopod. Extreme cases presented up to quadruplication of forelimbs (Fig. 4h and Extended Data Fig. 7b), a phenotype that is the inverse of *rspo2* crispant frogs that display total amelia (Extended Data Fig. 7a). We conclude that in the context of limb development, RSPO2 behaves as a direct antagonistic ligand to RNF43 and ZNRF3 without the need for LGR4/5/6. This ligand–receptor pair constitutes a master switch that governs the number of limbs an embryo should form. It will be important to assess whether this pathway can in part contribute to the disappearance of limbs during evolution, particularly in cetaceans and snakes, which are tetrapods that have become bi-amelic and tetra-amelic, respectively. It is also tempting to speculate whether the same

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

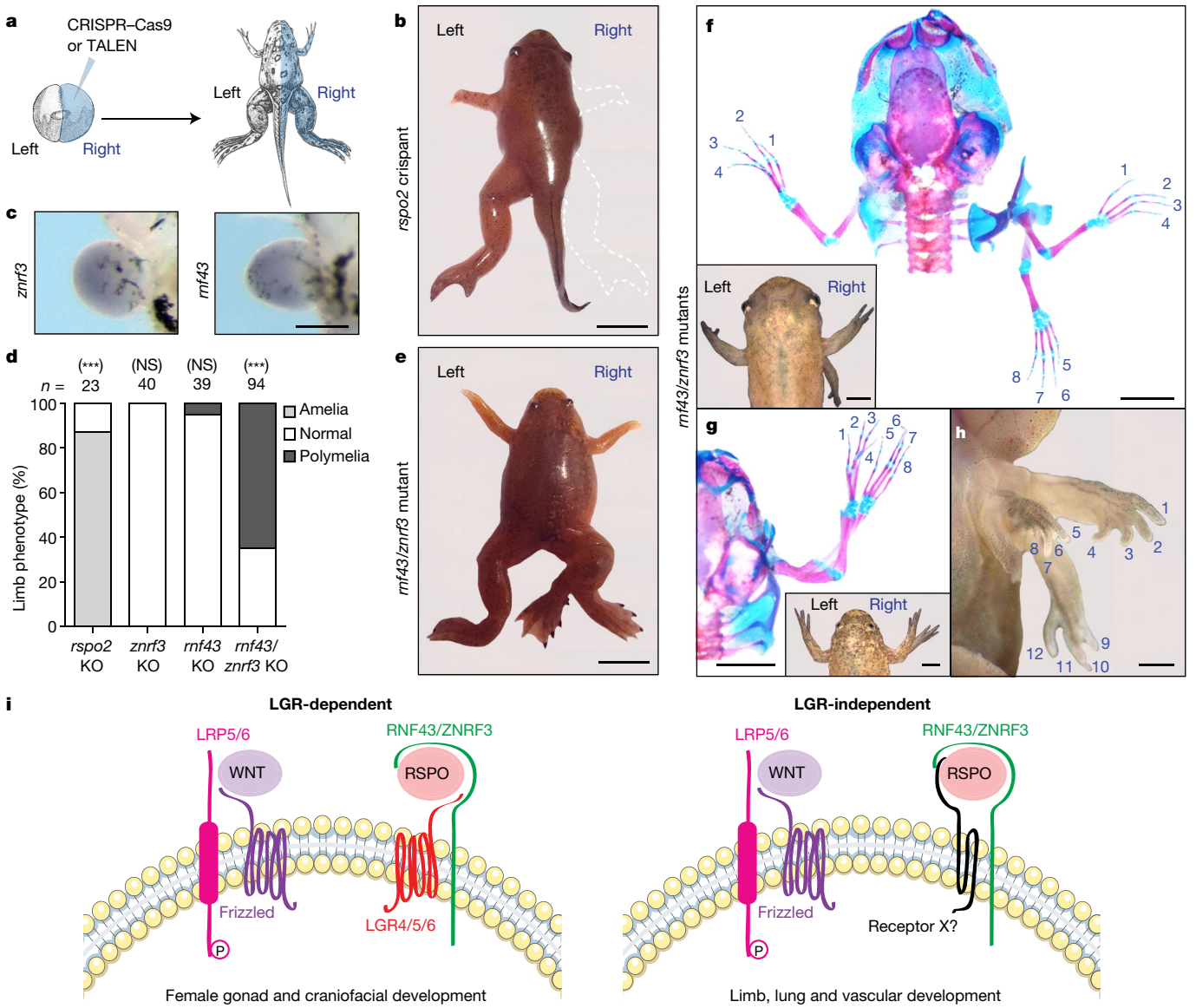


Fig. 4 | Frogs mutant for both *rnf43* and *znrf3* display complete limb duplications. a, Experimental scheme using *X. tropicalis*. **b**, Representative *rspo2* crispant (stage 63; $n = 21$). Scale bar, 0.5 cm. **c**, *znrf3* and *rnf43* in situ hybridization in stage 50 limb buds. Scale bar, 300 μm. The experiment was repeated three times. **d**, Scoring of limb phenotype. *n* denotes number of froglets. NS, not significant; $***P < 0.001$, significantly different from normal (χ^2 test). **e**, Stage 67 *znrf3/rnf43* double mutant with a duplicated right hindlimb. Scale bar, 0.5 cm. **f**, **g**, External view and alizarin red and alcian blue staining of *rnf43/znrf3* double mutants (stages 62 (**f**) and 66 (**g**)). Scale bars, 0.3 cm. **h**, *rnf43/znrf3* double-mutant tadpole (stage 59) displaying quadruplication of the right forelimb (three are visible). Scale bar, 0.2 cm. In **e**–**h**, 61 *znrf3/rnf43* double-mutant froglets with polymelia were obtained. **i**, Updated model for LGR-dependent R-spondin processes (left), and LGR-independent RSPO2/3 signalling (right), which may involve the activity of a hitherto unknown receptor X

embryonic signals may be re-mobilized in salamanders, which are capable of complete adult limb regeneration after amputation[30].

The current model suggests that RSPO–LGR form ligand–receptor pairs that serve to increase WNT signalling through direct inhibition of the two E3 ligases RNF43 and ZNRF3, which otherwise ubiquitinate WNT receptors for degradation[31,32]. Here we challenge this view and show that during embryogenesis, the concomitant loss of LGR4, LGR5 and LGR6 receptors does not phenocopy the loss of RSPO2 or RSPO3 (Fig. 4i).

Gain-of-function variants in *RSPO2* and *RSPO3* and loss-of-function alleles in *RNF43* and *ZNRF3* are the most frequent somatic mutations in colorectal cancer patients[33,34]. *LGR5*-positive cells have been shown to represent the major cell of origin of colorectal cancer[35]; however, pathogenic mutations in this WNT-associated receptor have not been documented so far. Our findings that RSPO2 and RSPO3 can inhibit RNF43 and ZNRF3, without the need for LGR4/5/6, raise the question of whether LGRs have any functional relevance to carcinogenesis. The ubiquitous triple-knockout *Lgr4/5/6* during embryogenesis serves as a proof-of-concept for subsequent organ-specific deletions, and should enable this question to be addressed.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0118-y.

Received: 9 November 2016; Accepted: 10 April 2018;
Published online: 16 May 2018

1. de Lau, W. et al. Lgr5 homologues associate with Wnt receptors and mediate R-spondin signalling. *Nature* **476**, 293–297 (2011).
2. Glinka, A. et al. LGR4 and LGR5 are R-spondin receptors mediating Wnt/β-catenin and Wnt/PCP signalling. *EMBO Rep.* **12**, 1055–1061 (2011).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

3. Carmon, K. S., Gong, X., Lin, Q., Thomas, A. & Liu, Q. R-spondins function as ligands of the orphan receptors LGR4 and LGR5 to regulate Wnt/β-catenin signaling. *Proc. Natl Acad. Sci. USA* **108**, 11452–11457 (2011).
4. Niswander, L. Pattern formation: old models out on a limb. *Nat. Rev. Genet.* **4**, 133–143 (2003).
5. Niemann, S. et al. Homozygous WNT3 mutation causes tetra-amelia in a large consanguineous family. *Am. J. Hum. Genet.* **74**, 558–563 (2004).
6. Kazanskaya, O. et al. R-Spondin2 is a secreted activator of Wnt/β-catenin signaling and is required for *Xenopus* myogenesis. *Dev. Cell* **7**, 525–534 (2004).
7. Kim, K. A. et al. R-Spondin proteins: a novel link to β-catenin activation. *Cell Cycle* **5**, 23–26 (2006).
8. Nam, J. S., Turcotte, T. J., Smith, P. F., Choi, S. & Yoon, J. K. Mouse cristin/R-spondin family proteins are novel ligands for the Frizzled 8 and LRP6 receptors and activate β-catenin-dependent gene expression. *J. Biol. Chem.* **281**, 13247–13257 (2006).
9. Ohkawara, B., Glinka, A. & Niehrs, C. Rspo3 binds syndecan 4 and induces Wnt/PCP signaling via clathrin-mediated endocytosis to promote morphogenesis. *Dev. Cell* **20**, 303–314 (2011).
10. Zebisch, M. et al. Structural and molecular basis of ZNRF3/RNF43 transmembrane ubiquitin ligase inhibition by the Wnt agonist R-spondin. *Nat. Commun.* **4**, 2787 (2013).
11. Bell, S. M. et al. R-spondin 2 is required for normal laryngeal-tracheal, lung and limb morphogenesis. *Development* **135**, 1049–1058 (2008).
12. Aoki, M., Kiyonari, H., Nakamura, H. & Okamoto, H. R-spondin2 expression in the apical ectodermal ridge is essential for outgrowth and patterning in mouse limb development. *Dev. Growth Differ.* **50**, 85–95 (2008).
13. Nam, J. S. et al. Mouse R-spondin2 is required for apical ectodermal ridge maintenance in the hindlimb. *Dev. Biol.* **311**, 124–135 (2007).
14. Yamada, W. et al. Craniofacial malformation in R-spondin2 knockout mice. *Biochem. Biophys. Res. Commun.* **381**, 453–458 (2009).
15. Brüchle, N. O. et al. RSPO4 is the major gene in autosomal-recessive anonychia and mutations cluster in the furin-like cysteine-rich domains of the Wnt signaling ligand R-spondin 4. *J. Invest. Dermatol.* **128**, 791–796 (2008).
16. Başaran, S. et al. Tetra-amelia, lung hypo-/aplasia, cleft lip-palate, and heart defect: a new syndrome? *Am. J. Med. Genet.* **51**, 77–80 (1994).
17. Sousa, S. B. et al. Tetra-amelia and lung hypo/aplasia syndrome: new case report and review. *Am. J. Med. Genet. A.* **146A**, 2799–2803 (2008).
18. Xie, Y. et al. Interaction with both ZNRF3 and LGR4 is required for the signalling activity of R-spondin. *EMBO Rep.* **14**, 1120–1126 (2013).
19. Proffitt, K. D. et al. Pharmacological inhibition of the Wnt acyltransferase PORCN prevents growth of WNT-driven mammary cancer. *Cancer Res.* **73**, 502–507 (2013).
20. Mazerbourg, S. et al. Leucine-rich repeat-containing, G protein-coupled receptor 4 null mice exhibit intrauterine growth retardation associated with embryonic and perinatal lethality. *Mol. Endocrinol.* **18**, 2241–2254 (2004).
21. Morita, H. et al. Neonatal lethality of LGR5 null mice is associated with ankyloglossia and gastrointestinal distension. *Mol. Cell. Biol.* **24**, 9736–9743 (2004).
22. Snippert, H. J. et al. Lgr6 marks stem cells in the hair follicle that generate all cell lineages of the skin. *Science* **327**, 1385–1389 (2010).
23. Styrkarsdottir, U. et al. Nonsense mutation in the *LGR4* gene is associated with several human diseases and other traits. *Nature* **497**, 517–520 (2013).
24. Sulem, P. et al. Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015).
25. Koizumi, M. et al. Lgr4 controls specialization of female gonads in mice. *Biol. Reprod.* **93**, 90 (2015).
26. Jin, Y. R., Turcotte, T. J., Crocker, A. L., Han, X. H. & Yoon, J. K. The canonical Wnt signaling activator, R-spondin2, regulates craniofacial patterning and morphogenesis within the branchial arch through ectodermal-mesenchymal interaction. *Dev. Biol.* **352**, 1–13 (2011).
27. Aoki, M. et al. R-spondin3 is required for mouse placental development. *Dev. Biol.* **301**, 218–226 (2007).
28. van de Glind, G. C. et al. RNA-seq analysis of Lgr6$^+$ stem cells and identification of an Lgr6 isoform. *Exp. Dermatol.* https://doi.org/10.1111/exd.13453 (2017).
29. Lebensohn, A. M. & Rohatgi, R. R-spondins can potentiate WNT signaling without LGRs. eLife 7, e33126 (2018).
30. Simon, A. & Tanaka, E. M. Limb regeneration. *Wiley Interdiscip. Rev. Dev. Biol.* **2**, 291–300 (2013).
31. Hao, H. X. et al. ZNRF3 promotes Wnt receptor turnover in an R-spondin-sensitive manner. *Nature* **485**, 195–200 (2012).
32. Koo, B. K. et al. Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors. *Nature* **488**, 665–669 (2012).
33. Seshagiri, S. et al. Recurrent R-spondin fusions in colon cancer. *Nature* **488**, 660–664 (2012).
34. Bond, C. E. et al. RNF43 and ZNRF3 are commonly altered in serrated pathway colorectal tumorigenesis. *Oncotarget* **7**, 70589–70600 (2016).
35. Barker, N. et al. Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**, 608–611 (2009).

Acknowledgements We are grateful to the families of affected fetuses whose cooperation made this study possible. We thank M. Krahn and D. Bertola for contributing more families with TETAMS for ongoing research. We thank C. Niehrs for the RSPO2-ΔC-AP-PCS2+construct. We thank D. Virshup for HEK293T-STF cells. The TALEN plasmids were obtained from C. H. K. Cheng. We are grateful to our groups' members for discussion and advice. E.S.-R. thanks M. Ravi for her patience and support, and is supported by a BMRC Young Investigator Grant (1613851035). S.S.N. is a recipient of Senior Research Fellowship from Indian Council of Medical Research (45/10/2015-HUM-BMS). N.B. is funded by A*STAR and an NRF Investigatorship. B.R. is a fellow of the Branco Weiss Foundation, an A*STAR Investigator, a NRF and AAA fellow and a Young EMBO Investigator. This work was partly funded by the Department of Science and Technology, India, for the project 'Application of autozygosity mapping and exome sequencing to identify genetic basis of disorders of skeletal development' (SB/SO/HS/005/2014); by the Research Foundation, Flanders (FWO-Vlaanderen) (grants G0A1515N and G029413N), the Concerted Research Actions from Ghent University (BOF15/GOA/01) and the Hercules Foundation, Flanders (grant AUGE/11/14), Belgium; by TUBITAK 108S412 and 112S398 (ERA-Net consortiums, CRANIRARE and CRANIRARE2), Turkey; and by a Strategic Positioning Fund on Genetic Orphan Diseases from A*STAR, Singapore.

Reviewer information *Nature* thanks C. Niehrs, C. Tabin and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions The recruitment of families 1 and 2 and the initial discovery of *RSPO2* mutations were made by the team of H.K. with the help of U.A. and Z.O.U. Further patient data and samples were collected by H.K. from S.B.d.S., S.S.N., A.S. and K.M.G., who provided TETAMS families 3 and 5. C.B., X.L. and C.L.C. independently identified a *RSPO2* mutation in family 4. E.S.-R., M.L., K.V., N.S. and B.R. designed functional studies. N.B. provided single *Lgr4*, *Lgr5* and *Lgr6* knock-in mouse lines. E.S.-R., U.A., C.B., B.K., X.L., S.B., E.B.Y., C.L.C. and Z.O.U. performed whole-exome sequencing, chromosomal array analysis and sequencing analysis. E.S.-R., M.L., C.B.-L., M.K., H.T.T., T.N., R.N., A.H., N.S., T.T.T. and L.V. performed functional experiments. C.B.-L., M.K. and H.T.T. contributed equally to this work. E.S.-R., U.A., M.L., K.V., H.K. and B.R. wrote and revised the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0118-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0118-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to K.V. or N.B. or H.K. or B.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Fetuses and clinical assessment. The five families included in this study were enrolled from genetic departments of Istanbul, Turkey; Coimbra, Portugal; Nantes, France; and Mangalore, India. Eight out of eleven affected individuals had been clinically and radiologically evaluated by experienced clinical geneticists. Autopsy was performed in five cases (see Extended Data Table 1). Families 2 and 3 had been previously reported[16,17]. Written informed consent in accordance with the Helsinki protocol was obtained from family members before inclusion to the research protocol. Consent to publish photos was obtained from the families. The studies were performed in compliance with all relevant ethical regulations from the respective institutions. Approvals were obtained from the Istanbul University, Istanbul Medical Faculty ethical committee, Turkey (CRANIRARE: 2008/1194 and CRANIRARE-II: 2012/743-1061) as well as the Koç University School of Medicine (KUSoM) ethical committee, Istanbul, Turkey (2015.120.IRB2.047 CRANIRARE-2) for families 1 and 2; from the 'Comissão de Ética do Centro Hospitalar de Coimbra', Coimbra, Portugal (2009/42, 1724/Sec) for family 3; from the 'Comite de protection des personnes Ouest IV', Nantes, France (DC-2011-1399) for family 4; and from the institutional ethics committee of Kasturba Hospital, Manipal, India (ECR/146/Inst/KA/2013, IEC 430/2013) for family 5.

Genotyping and exome sequencing. DNA was extracted from skin biopsy samples of affected cases and from peripheral blood leukocytes of parents and healthy siblings by standard procedures. Affected individuals, or parents in the absence of samples from the affected individuals, were previously screened and excluded for any functional sequence variations/mutations in the *WNT3* (NM_030753) and *WNT7A* (NM_004625) genes. Whole-exome capture of subjects II:6 and II:7 from family 1 were performed using Agilent SureSelect Human All Exon v4.0 kit, sequenced on Illumina HiSeq2000 platform using TruSeq v3 chemistry at a mean coverage of 50 ×, reads provided in Fastq files were mapped to human genome (hg19) using Burrows–Wheeler Aligner (BWA package version 0.6.2). Local realignment was performed by Genome Analysis Tool Kit (GATK). Duplicated reads were marked for exclusion from further analysis using Picard (version 1.83) tool. Further alignment manipulations were performed by Samtools (version 0.1.18). Base quality (Phred scale) scores were recalibrated using the GATK covariance recalibration for each sample (Oxford Gene Technology), and variant calling was performed using ANNOVAR tool with avSNP release of 142, 1,000 genomes release of 2014 along with NIH-NHLBI 6500 exome database version 2. All the alterations including overlapping homozygous variants complying with minimum of 20 read depths were considered for further evaluation and browsed on OGT NGS (Oxford Gene Technologies' Next Generation Sequencing) and IGV (Integrated Genomic Viewer) browsers. Screening for *RSPO2* mutations in additional affected individuals or parents, and segregation validation were performed by Sanger sequencing, with PCR primers designed to cover all the coding exons and the flanking regions according to RefSeq accession number NM_178565 (Extended Data Table 2a).

Array-CGH and SNP-array analysis. Oligonucleotide array-CGH was performed using SurePrint G3 Human CGH Microarray ISCA 4 × 180 K v2 (Agilent Technologies), and the SNP-Array adopted was 300 K HumanCytoSNP-12v2-1 (Illumina Inc.). The 180 K kit has an overall median probe spacing of 13 kb, and the SNP-Array has 6.2 kb. Analyses were performed according to the protocols provided by the suppliers (Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis and Illumina Karyostudio & Bluefuse Multi Softwares). Arrays were scanned using a NimbleGen MS 200 for Agilent SurePrint array and I-Scan instrument for HumanCytoSNP-12v2-1. Genomic positions were based on the UCSC February 2009 human reference sequence (hg19) (NCBI build 37 reference sequence assembly).

Constructs. An RSPO2-ΔC-AP-pCDNA3 plasmid (gift from C. Niehrs) encoding the wild-type human *RSPO2* open-reading frame (ORF) without the C-terminal domain (ΔC) (NP_848660 amino acids 1–206) tagged with alkaline phosphatase (AP) was used to generate the RSPO2-ΔC-AP R69C and F105A/F109A mutant constructs (RSPO2(R69C) and RSPO2(F105A/F109A)) using the QuikChange Mutagenesis Kit (Stratagene 200522). R69C is the missense mutation found in family 1, and the F105A/F109A mutations specifically abolish the interaction of RSPO2 with the LGRs[36,37]. RSPO2-Δ70-AP (RSPO2(Q70X), mutation found in family 2) was obtained by PCR and ligation (deleting the C-terminal region of the protein downstream of position 69). Deletion of the RSPO2 C-terminal domain (RSPO2-ΔC) decreases its retention on the cell surface without affecting its receptor binding and WNT enhancement properties[38]. A construct for a secreted alkaline phosphatase was used as a negative control. For cell surface binding assay experiments, V5-LGR5-pCS2 + (gift from C. Niehrs[2]), and pCMV6-Entry-RNF43 (Origene RC214013) plasmids were used. For co-immunoprecipitation experiments, the signal peptide and extracellular domains (ECD) of LGR5 (NP_001264156 amino acids 1–557), ZNRF3 (NP_001193927 amino acids 1–219), and RNF43 (NP_060233 amino acids 1–197) were subcloned in pCS2 + with a Flag tag at their C terminus.

HEK293T and HEK293T-STF cell culture. HEK293T (from ATCC) and HEK293T-STF (SUPERTOPFLASH, gift from D. Virshup, from ATCC) cell lines have not been authenticated but were tested negative for mycoplasma contamination. They were cultured on plates coated with poly-L-lysine (Sigma P4707) with the following medium: DMEM High glucose (HyClone SH30081.01) with 10% fetal bovine serum (Thermo Scientific SH30070), and 2 mM L-glutamine (ThermoFisher Scientific 25030081). Cells were transfected with DNA plasmid using the FuGENE HD Transfection Reagent (Promega E2312) in OptiMEM medium (Gibco 31985070).

Expression and secretion studies. For protein extraction, HEK293T cells were lysed using appropriate amount of RIPA buffer (Tris-HCl pH7.5, 50 mM, NaCl 150 mM, NP-40 0.1%, Na^{2+}-deoxycholate 0.05%) supplemented with proteinase inhibitors (Complete, Roche 04693159001). Lysates were centrifuged at 17,000*g* for 15 min at 4 °C to remove cell debris, and the supernatants (protein extracts) were collected. For secretion studies, culture medium was changed 24 h after transfection with a serum-free medium Pro293a-CDM (Lonza 12-764Q) supplemented with L-glutamine without or with 50 μg ml^{-1} of heparin (Sigma H3149). Secretion was allowed for 48 h before collection of the conditioned media. For western blotting, samples were electrophoresed with reducing Laemeli loading buffer after denaturation at 95 °C for 10 min. The protein ladder (Bio-Rad 161-0377) and denatured and reduced samples were loaded onto 4–20% gradient precast gels (BioRad Criterion 567-1093) in 1 × running buffer (25 mM Tris, 200 mM glycine, 0.1% SDS) and ran at 80–180 V until desired separation. Gels were transferred onto 0.2 μm PVDF membranes (BioRad Criterion 170-4157) using the Trans-Blot TurboTM transfer system for 7 min. Membranes were blocked for 1 h at room temperature with 5% milk in TBST (50 mM Tris-HCl pH 7.5, 150 mM NaCl; with 0.05% Tween20). Membranes were incubated with primary antibody diluted in 5% milk in TBST at 4 °C overnight (anti-alkaline phosphatase, 1:2,000, GenHunter Q310; anti-Flag, 1:1,000, Cell Signaling 14793 S; anti-GAPDH, 1:4,000, SantaCruz 47724). After washes in TBST, membranes were incubated for 1 h at room temperature with secondary antibodies (Mouse-HRP 71503510 or Rabbit-HRP 711035152, 1:4,000, Jackson Immuno) in 5% milk in TBST. After several washes in TBST, the signal was revealed with the HRP substrate (Thermo Scientific SuperSignal 34080/34076/34096) for 3 min at room temperature. Membranes were then exposed to CL-Xposure films (Thermo Scientific 34091), and developed in a Carestream Kodak developer.

Co-immunoprecipitation experiments. Conditioned media containing either of the RSPO2-ΔC-AP forms or the different receptor-ECD-Flag proteins were obtained after transfection in HEK293T. Conditioned media with equivalent amount of each RSPO2-ΔC-AP forms were first mixed with conditioned medium containing the receptor-ECD-Flag of interest for 4 h at 4 °C. At the same time, Protein G Dynabeads (Novex 10003D) were conjugated with anti-Flag antibodies (Sigma F3165) for 4 h at 4 °C. The media mixes (inputs) were then incubated with the conjugated beads overnight at 4 °C. After washes, the beads were re-suspended with 2 × reducing Laemeli loading buffer. After centrifugation, the supernatants (immunoprecipitants) were subsequently used for western blotting.

Cell-surface binding assay. Twenty-four hours after transfection of HEK293T cells with V5-LGR5-pCS2 + (gift from C. Niehrs[2]), pCMV6-Entry-RNF43 (Origene RC214013) or pCS2 + (empty vector) in 24-well plates, the cell culture medium was replaced for 3 h with 300 μl of conditioned medium containing equivalent amounts of RSPO2-ΔC-AP proteins (as determined by western blot), to assess for their cell-surface binding. After washes with PBS, cells were lysed with PBS containing Triton X-100 1% and 1 × proteinase inhibitor, and then incubated at 65 °C for 1 h to inhibit endogenous alkaline phosphatase activity. After centrifugation at 17,000 *g* for 2 min, supernatants (protein extracts) were collected and the protein concentration was measured (Pierce BCA protein assay kit 23225). Eighty microlitres of the same quantity of total proteins for each condition was added to 80 μl of BM Purple (Roche 11442074001) and incubated overnight at 4 °C in the dark for chromogenic development. Pictures were taken with the NCS Microtek Artixscan F1 scanner.

SUPERTOPFLASH luciferase assay. HEK293T-STF cells were transfected with the human *WNT3A* gene (hWNT3A-pCS2 +) and the *Renilla* luciferase (pRL-CMV vector). RSPO2-ΔC-AP constructs were either transfected or conditioned media was added 24 h after transfection for another 24 h incubation. The expression of the firefly (STF) and *Renilla* luciferases were measured using the Dual-Luciferase Reporter Assay system (Promega E1960) 48 h after transfection. Measurements were done on opaque 96-well plates using a luminometer. Luminescence data are represented as the firefly luminescence relative to the *Renilla* luminescence and total protein concentration. Plotted are the values relative to the values for WNT3A plus alkaline phosphatase alone. For SV40-immortalized mouse fibroblasts-STF, cells were treated for 24 h with recombinant proteins re-suspended in PBS containing 0.1% BSA. Luminescence data are represented as the firefly luminescence relative to total protein concentration. Plotted are the values relative to the values for WNT3A alone. Data are average of three independent experiments and

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

statistical analyses were done using the PRISM5 software using a one-way ANOVA test with Bonferroni correction for multiple hypothesis testing. Not significant indicates $P > 0.05$. $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$. Error bars indicate s.e.m.

Mouse lines. All experiments on mice were executed in accordance with the guidelines and regulations of the respective institutions. Approvals were obtained in Singapore by the Institutional Animal Care and Use Committee (IACUC 140907, 171263, and 171210) and in Germany by the ethical committee of the Institute of Molecular Biology (IMB) gGmbH (permit number 23-177-07/G 16-05-071E1). *Lgr4-eGFP-ires-creERT2, Lgr5-eGFP-ires-creERT2* and *Lgr6-eGFP-ires-creERT2* mice have been described previously[1,22,39]. We incrossed $Lgr4^{+/-}Lgr5^{+/-}Lgr6^{-/-}$ animals and analysed 82 embryos (63 embryos at E14.5, and 19 embryos at E18.5, both male and female). See Extended Data Table 2b for the genotyping PCR primers used. For Porcupine inhibitor treatments, Wnt-C59 (Tocris, 5148/10) or vehicle (PBS containing 0.5% methylcellulose, 0.01% Tween-20, and 5% DMSO) was orally administered to wild-type pregnant females at 5 μg g^{-1} dam body weight every 12 h between E9.75 and E14.25, and embryos were collected at E14.5. We analysed a total of 30 embryos from vehicle-treated mothers, and 46 embryos from Wnt-C59-treated mothers. No statistical method was used to predetermine sample size. No randomization or blinding was used during the experimental procedures.

Mouse embryos sections. Embryos were fixed overnight at 4 °C in 4% paraformaldehyde and stored in 70% ethanol at 4 °C. For H&E staining, RNAscope and immunofluorescence experiments, embryos were subsequently paraffin-embedded and sectioned using a Leica RM2255 microtome. H&E staining was performed on deparaffinized and rehydrated 6 μm tissue sections according to standard protocols, and images were taken with the MetaSystems Metafer Slide Scanner. RNAscope experiments were performed on deparaffinized and rehydrated 6 μm tissue sections using the RNAscope 2.5 HD Duplex Assay (Advanced Cell Diagnostics ACD 322430). The ACD RNAscope probes used in this study are as follow: Mm-Rspo2-C2 (402001-C2), Mm-Lgr4 (318321), Mm-Lgr5 (312171), Mm-Lgr6 (404961), Mm-Rnf43 (400371), Mm-Znrf3 (434201), Mm-Wnt3 (312241). RNAscope images were taken using a Zeiss AxioImager Z1 upright microscope with the ZEN software. Immunofluorescence was performed on deparaffinized and rehydrated 7 μm tissue sections. Antigen retrieval was carried out by heating slides in a pressure cooker (121 °C) for 20 min at pH 9.0 (S2367, DAKO). The following primary antibodies were used: chicken anti-eGFP (1:2,000, Abcam ab290), rabbit anti-vimentin (1:500, Abcam ab92547), mouse anti-E-cadherin (1:200, BD Transduction Laboratory 61018). The secondary antibodies used were: anti-chicken/rabbit/mouse Alexa 488/568/647 IgG (1:500, Invitrogen). Images were taken on the Olympus FV3000 inverted confocal microscope.

Generation of mouse cell lines. Embryos were collected at E14.5 and primary dermal fibroblasts were derived from embryonic back skin explants, and primary NPCs were isolated from brain cortices. Primary fibroblasts were immortalized via SV40 large T antigen retroviral infection using standard protocol (pBABE-neo largeTcDNA, Addgene plasmid 1780). SV40-immortalized mouse fibroblasts-SUPERTOPFLASH (STF) were generated with the insertion of 7xTcf-firefly luciferase by retroviral infection using standard protocol (7TFP, Addgene plasmid 24308). For NPC derivation, cortices were first excised from whole brains and mechanically dissociated using a pipette. Cells were cultured as suspension in NPC medium: DMEM/F12 (Gibco 11320) containing 1 × N-2 supplement (ThermoFisher Scientific 17502001), 1 × B-27 supplement (ThermoFisher Scientific 17504001), 2 mM L-glutamine (ThermoFisher Scientific 25030081), 0.1 mM NEAA (Gibco 1140-050), 20 ng ml^{-1} FGF-2 (R&D Systems 233-FB) and 20 ng ml^{-1} EGF (R&D Systems 236-eg). NPCs in suspension were passaged with accutase (Merck Millipore SCR005) every 4–5 days. For adherent cultures, NPCs were plated onto Matrigel-coated (Corning 354231) plates in NPC medium. Adherent mouse NPCs were reprogrammed by transduction of human *OCT4* (also known as *POU5F1*), *SOX2*, *KLF4* and *MYC* (Addgene plasmids 17225, 17226, 17227 and 18119)[37] using standard protocol. After 4 days, transduced cells were reseeded onto irradiated mouse embryonic fibroblasts in mouse iPS cell medium: knockout-DMEM (ThermoFisher Scientific 1029018) supplemented with 10% Knock Out Serum Replacement (ThermoFisher Scientific 10828028), 10% FBS (ThermoFisher Scientific 16000044), 0.1 mM 2-mercaptoethanol (ThermoFisher Scientific 21985023), 2 mM L-glutamine (ThermoFisher Scientific 25030081), 0.2 mM NEAA (ThermoFisher Scientific 1114050) and 1,000 U ml^{-1} mouse LIF (Stem Cell Technologies 78056). iPS cell colonies were picked between days 7 and 15 and maintained in mouse iPS cell medium for expansion on gelatin-coated plates.

Cell treatment with recombinant proteins. For WNT signalling response experiments, immortalized fibroblasts were treated for 24 h in serum-free medium with 25 ng ml^{-1} recombinant WNT3A (R&D Systems 5036-WN), and/or 400 ng ml^{-1} recombinant RSPO1 (4645-RS), RSPO2 (3266-RS), RSPO3 (3500-RS) or RSPO4 (4575-RS) re-suspended in PBS containing 0.1% BSA.

siRNA experiments. siGENOME SMARTpool mouse *Rspo2* (Dharmacon 239405), *Rspo3* (72780), *Znrf3* (407821) or negative control (D-001206-14-05) siRNAs were used to transfect SV40-immortalized mouse fibroblasts. The transfection of 37.5 nM siRNA was performed with Lipofectamine RNAiMAX (Invitrogen 13778-075) or DharmaFECT (Dharmacon T-2001) transfection reagents according to manufacturer protocols.

qPCR experiments. For qPCR experiments, embryonic organs or culture cells were lysed in the QIAGEN RLT buffer and total RNAs were extracted using the QIAGEN RNeasy Mini kit (74106), including the optional DNase RNase-free treatment. cDNAs were obtained using the iScript reverse transcription supermix (Bio-Rad 170-8841). qPCR were performed with primers described in Extended Data Table 2c using the Power SYBR Green Master mix (Applied Biosystems 4367659) on the Applied Biosystems 7900HT Fast Real-Time PCR system. Plotted are data relative to *Actb* and to the control condition. Data are the average of at least three biological triplicates and statistical analyses were done with the PRISM5 software using a one-way ANOVA test with Bonferroni correction for multiple hypothesis testing when more than two groups were compared, or an unpaired *t*-test with Welch's correction when less than three groups were compared.

Xenopus tropicalis experiments. All experiments on *X. tropicalis* were executed in accordance with the guidelines and regulations of Ghent University, faculty of Sciences, Belgium. Approval was obtained by the ethical committee of Ghent University, faculty of Sciences (permit number EC2017-093). No statistical method was used to predetermine sample size. No randomization or blinding was used during the experimental procedures.

***Xenopus tropicalis* whole-mount in situ hybridization.** Probes for *rnf43* and *znrf3* were designed by amplifying the coding sequence by PCR with primers linked to RNA-polymerase sites (Supplementary Table 1). Sense and antisense RNA probes were generated by in vitro transcription with the appropriate RNA polymerase and digoxigenin-rUTP-label. Whole-mount in situ hybridization was carried out as previously described[40]. Imaging was performed with a Leica MZ FLIII stereomicroscope/Leica DC300F camera.

Generation of *X. tropicalis* mosaic mutants by TALEN or CRISPR–Cas9. TALENs were generated using the Golden Gate Cloning protocol as previously described[41] and yields were quantified by Nanodrop (ThermoFischer Scientific). Embryos (two-cell stage) were injected unilaterally with 75 pg rnf43-TALEN-ELD, rnf43-TALEN-KKR, znrf3-TALEN-ELD and znrf3-TALEN-KKD (Extended Data Table 2e). Guide RNA targeting *rspo2* (Extended Data Table 2e) was designed with CRISPRScan (http://www.crisprscan.org/) and generated as previously described[42,43], and yield was quantified by Qubit BR RNA assay (ThermoFischer Scientific). Embryos were injected unilaterally at the two-cell stage with 47 pg *rspo2* gRNA and 900 pg NLS-Cas9-NLS (VIB/UGent Protein Service Facility). Both male and female animals were included in the study once they passed Nieuwkoop stage 59 and were scored as normal, or displaying either amelia (absence of at least one limb) or polymelia (at least one limb showing signs of duplication) by stereomicroscopic examination of limbs. For quantitative analysis of genome editing, nine embryos per injected clutch were pooled and lysed overnight in lysis buffer (50 mM Tris pH 8.8, 1 mM EDTA, 0.5% Tween-20, 200 μg ml^{-1} proteinase K) at 55 °C. Genotyping PCRs were performed with the respective primer pairs shown in Extended Data Table 2f. Targeted deep sequencing of amplicons was performed as previously described and analysed by the BATCH-GE software package[44]. Indel frequency data and sequence variants for all targeted deep sequencing are shown in Supplementary Table 1.

Skeletal staining of *X. tropicalis* tadpoles and froglets. For staining of the skeleton of the mutant animals, premetamorphic tadpoles and postmetamorphic froglets were euthanized using a 0.05% benzocaine solution. Animals were fully eviscerated, skinned and eyes were removed. Whole-mount alcian blue and alizarin red staining was performed as follows: 95% ethanol (4 days, change after 24 h), 100% acetone (48 h), 0.15% alcian blue 8GX (Sigma-Aldrich A3157) in 76% ethanol/20% glacial acetic acid/4% H_2O (24 h), 70% ethanol (24 h), 95% ethanol (12 h), 1% KOH (6 h), 0.05% alizarin red S (Sigma-Aldrich A5533) in 1% KOH (48 h), 1% KOH (48 h). Imaging was performed with a Leica MZ FLIII stereomicroscope/Leica DC300F camera.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

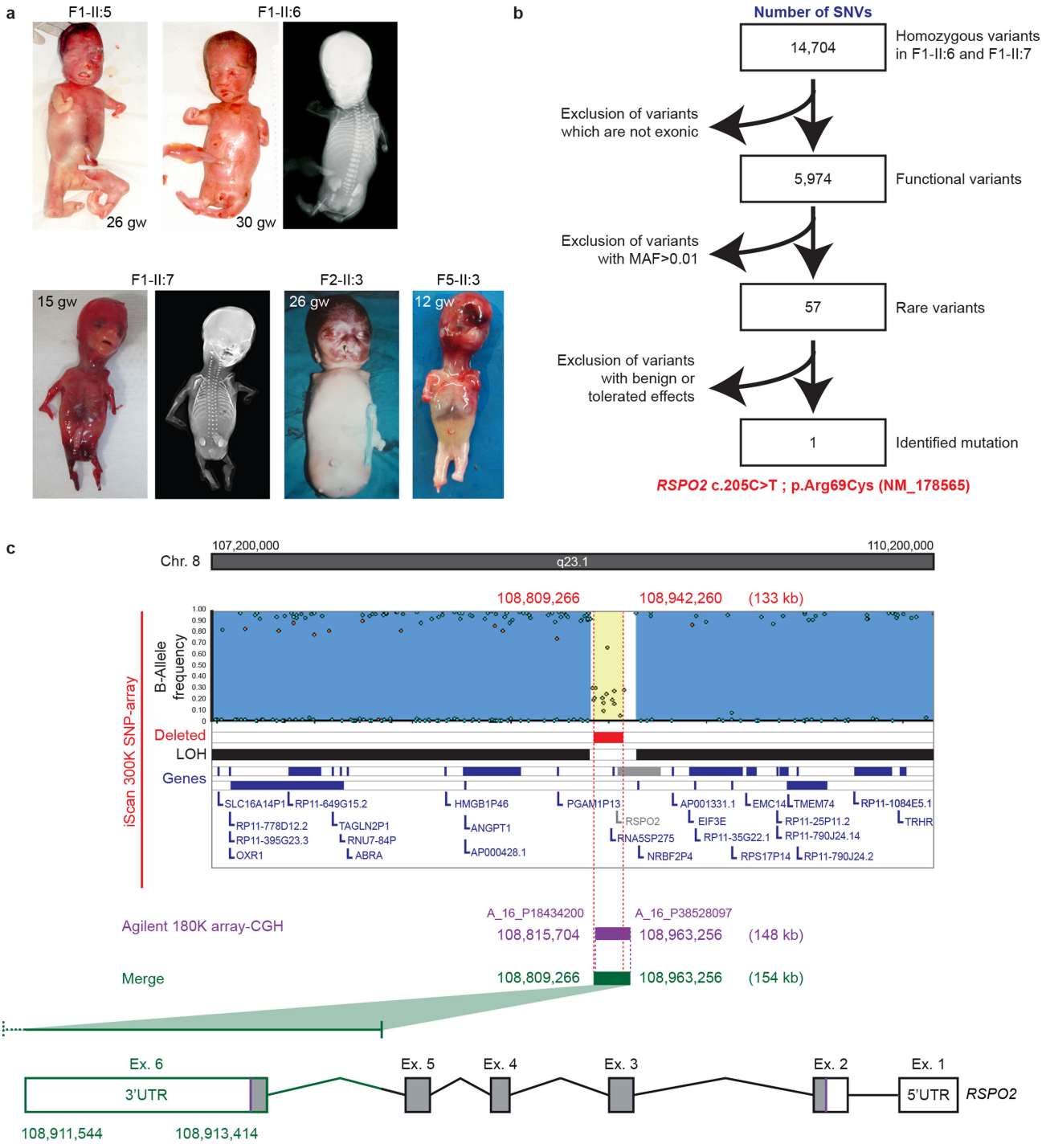
Data availability. The data that support the findings of this study are available within the paper or from the corresponding authors upon reasonable request. The whole exome sequencing data for family 1, and the SNP-array data for F3-II:1 shown in Extended Data Fig. 1 have been deposited in the NCBI Gene Expression Omnibus (GEO) database under accession numbers SRP136052 and GSE111781, respectively.

36. Wang, D. et al. Structural basis for R-spondin recognition by LGR4/5/6 receptors. *Genes Dev.* **27**, 1339–1344 (2013).
37. Peng, W. C. et al. Structure of stem cell growth factor R-spondin 1 in complex with the ectodomain of its receptor LGR5. *Cell Reports* **3**, 1885–1892 (2013).
38. Li, S. J. et al. Loss-of-function point mutations and two-furin domain derivatives provide insights about R-spondin2 structure and function. *Cell. Signal.* **21**, 916–925 (2009).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

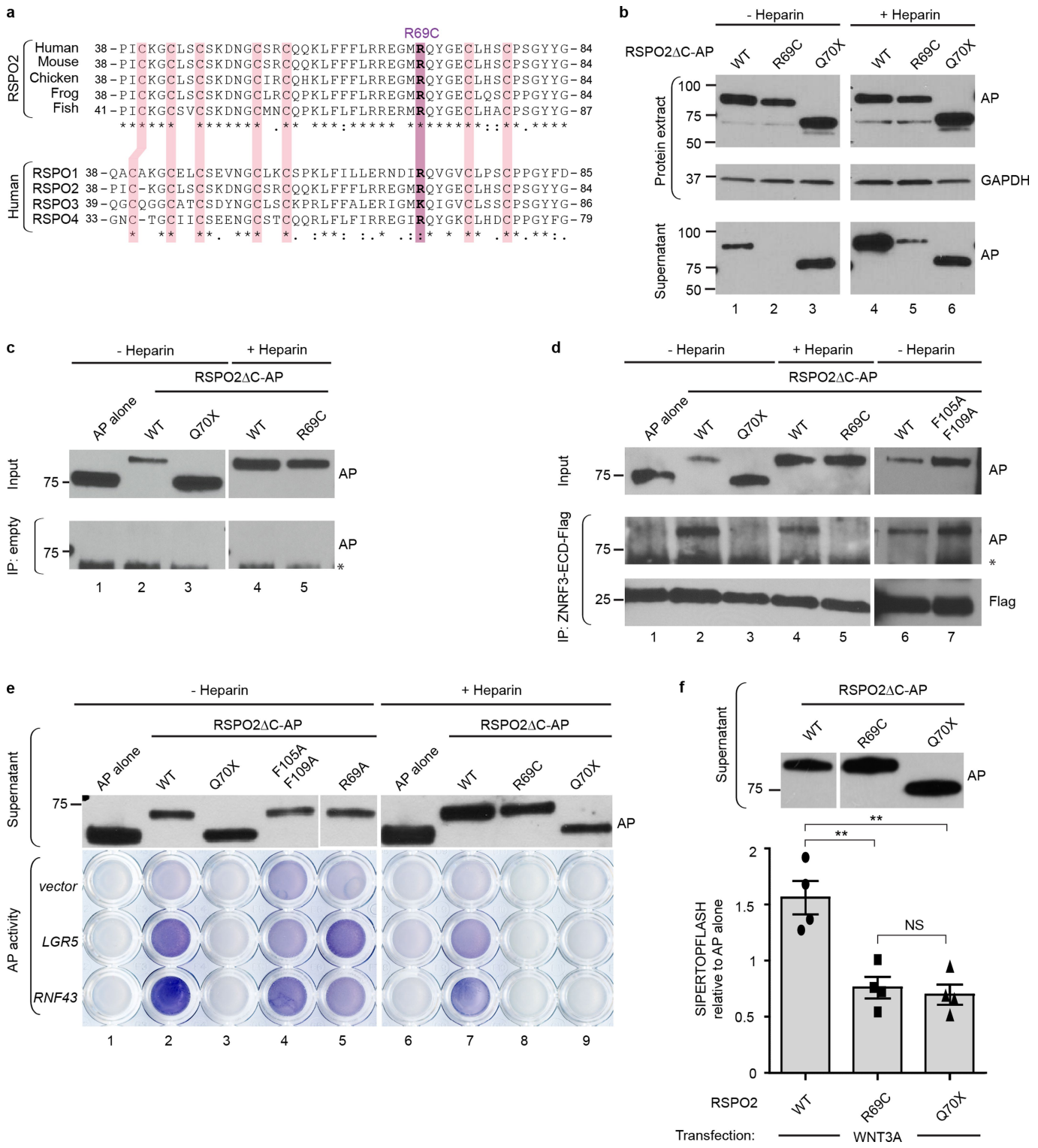
39. Krug, U., Patzschke, R., Zebisch, M., Balbach, J. & Sträter, N. Contribution of the two domains of *E. coli* 5′-nucleotidase to substrate specificity and catalysis. *FEBS Lett.* **587**, 460–466 (2013).
40. Monsoro-Burq, A. H. A rapid protocol for whole-mount in situ hybridization on *Xenopus* embryos. *CSH Protoc.* **2007**, pdb.prot4809 (2007).
41. Van Nieuwenhuysen, T. et al. TALEN-mediated apc mutation in *Xenopus tropicalis* phenocopies familial adenomatous polyposis. *Oncoscience* **2**, 555–566 (2015).
42. Naert, T., Van Nieuwenhuysen, T. & Vleminckx, K. TALENs and CRISPR/Cas9 fuel genetically engineered clinically relevant *Xenopus tropicalis* tumor models. *Genesis* **55**, https://doi.org/10.1002/dvg.23005 (2017).
43. Moreno-Mateos, M. A. et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12**, 982–988 (2015).
44. Boel, A. et al. BATCH-GE: Batch analysis of Next-Generation Sequencing data for genome editing assessment. *Sci. Rep.* **6**, 30330 (2016).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 1 | Pictures of affected fetuses, exome sequencing analysis in family 1 and genotyping analysis in family 3. a, Pictures and radiographs of indicated fetal cases illustrating severe limb defects in fetuses with HFH-RTRD, and the complete absence of limbs in fetuses with TETAMS. **b**, Summary of exome sequencing analysis for family 1, revealing a single biallelic missense mutation in the *RSPO2* gene. **c**, Summary of genotyping in family 3 by SNP-array and array-CGH analysis, revealing a homozygous deletion including exon 6 of *RSPO2*.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 2 | The RSPO2(R69C) and RSPO2(Q70X) mutants fail to bind ZNRF3. a, The RSPO2 R69 residue (highlighted in purple) is highly conserved in vertebrates and within the human paralogues RSPO1–RSPO4. Conserved cysteine residues of the Furin-like 1 domain are highlighted in pink. Protein alignment performed with ClustalO. **b**, Western blotting of protein extracts and supernatants from HEK293T cells transfected with indicated constructs. Deletion of the RSPO2 C-terminal domain (RSPO2-ΔC) decreases its retention on the cell surface without affecting its receptor binding and WNT enhancement properties[38]. The RSPO2(R69C) mutant was almost undetectable in conditioned media but was greatly increased by the addition of heparin in the medium. **c**, Co-immunoprecipitation of wild-type and mutant forms of RSPO2-ΔC-AP with the ProteinG-Flag beads only. Asterisk indicates an unspecific band. **d**, Co-immunoprecipitation of wild-type and mutant forms of RSPO2-ΔC-AP with the ZNRF3-ECD-Flag E3 ligase. Asterisk indicates an unspecific band. **e**, Cell-surface binding assay of HEK293T cells transfected with empty vector, *LGR5* or *RNF43*, using equivalent amounts of RSPO2-ΔC-AP conditioned media (western blot). Experiments in **b**–**e** were repeated three times. **f**, SUPERTOPFLASH assay in HEK293T-STF cells transfected with WNT3A in the presence of equivalent amounts of RSPO2-ΔC-AP conditioned media (western blot). $n = 4$ biological replicates. Data are mean ± s.e.m. NS, not significant. $^{**}P < 0.01$, one-way ANOVA test with Bonferroni's correction. For gel source data, see Supplementary Fig. 1.

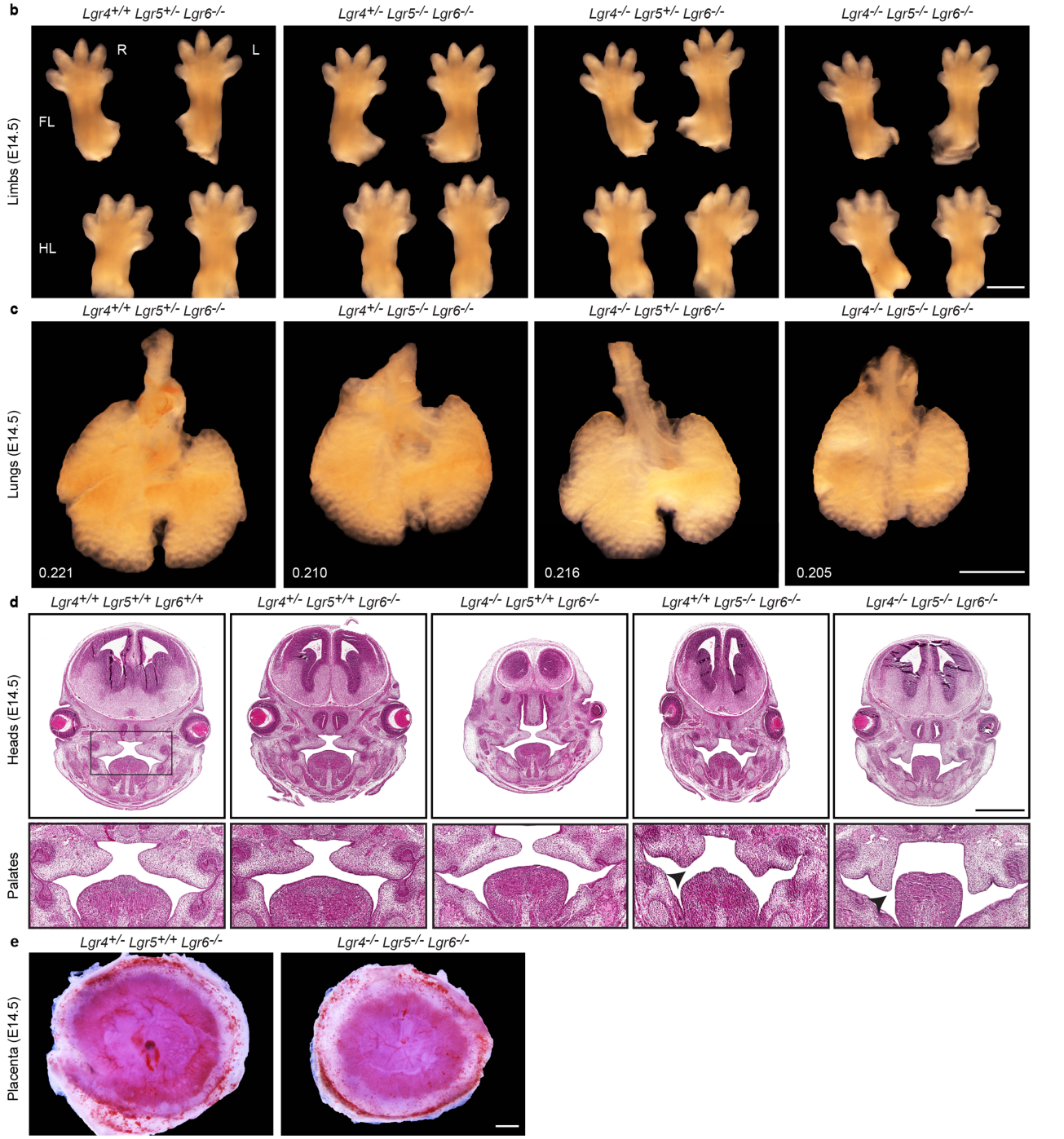
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

a

Lgr4$^{+/-}$ *Lgr5*$^{+/-}$ *Lgr6*$^{-/-}$ incrosses

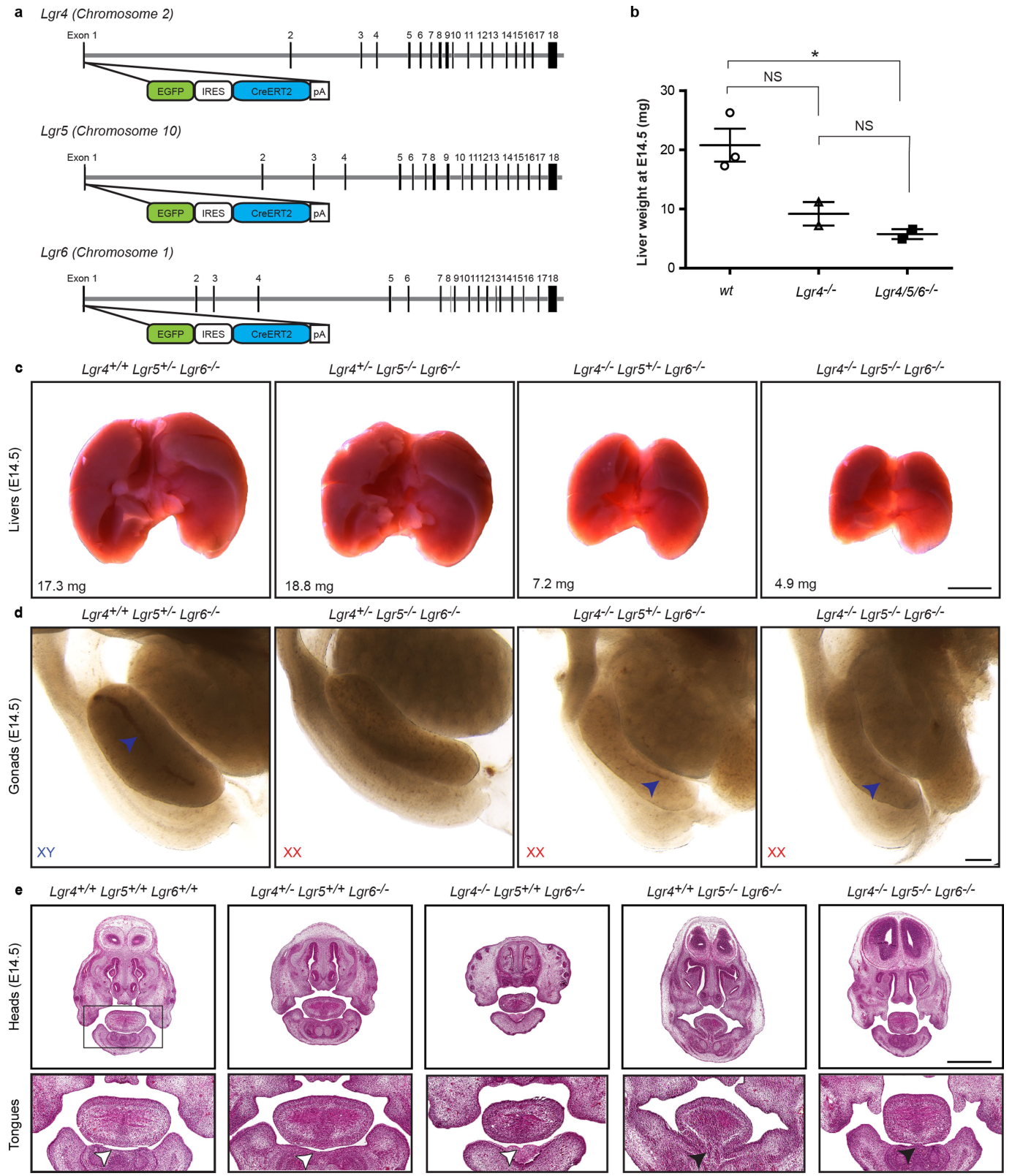
Embryos	Expected		Obtained Lgr4		Obtained Lgr5	
n=82	n	%	n	%	n	%
wt	20.5	25	19	23.2	20	24.4
het	41	50	48	58.5	37	45.1
hom	20.5	25	15	18.3	25	30.5

Triple hom	Expected	Obtained
82/16 =	5.125	5



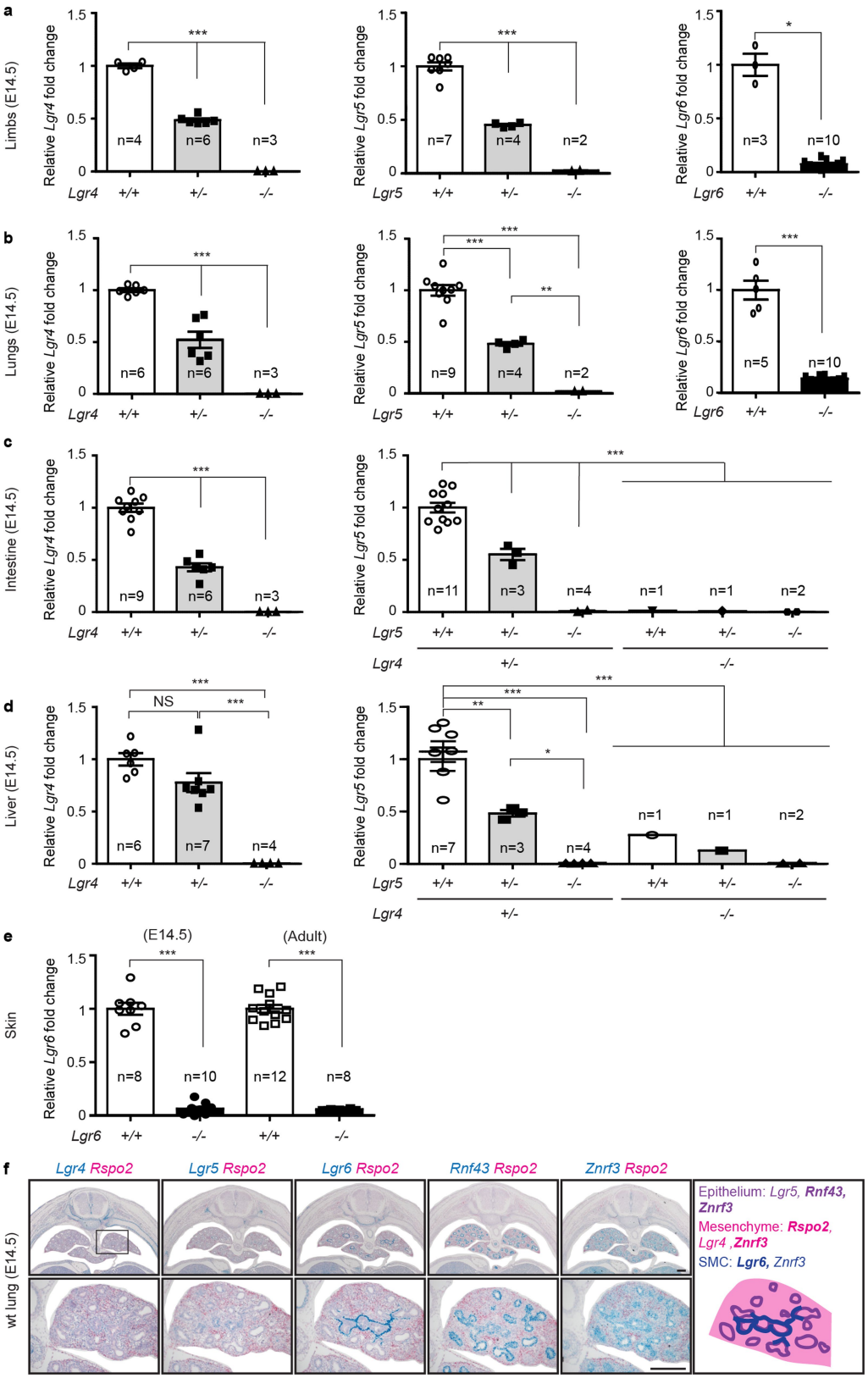
Extended Data Fig. 3 | Mouse *Lgr4/5/6* knock-in embryos do not recapitulate the *Rspo2* and *Rspo3* phenotypes. a, Table indicating the proportion and numbers of analysed embryos. The different genotypes were obtained in Mendelian ratio ($P > 0.05$, χ^2 test). **b**, Normal limbs of embryos with indicated genotypes including a *Lgr4/5/6* triple-knockout embryo. All obtained embryos (see **a**) had a similiar phenotype. FL, forelimbs; HL, hindlimbs; L, left; R, right. Scale bar, 1 mm. **c**, Normal lungs of embryos with indicated genotypes including a *Lgr4/5/6* triple-knockout embryo. Comparable lung length relative to body length is indicated. Scale bar, 1 mm. All obtained embryos (see **a**) had a similiar phenotype. **d**, H&E staining of coronal sections of the heads through the oral cavity. Scale bar, 1 mm. Close-ups of the palatal shelves illustrate the cleft palate present in the *Lgr5/6* double-knockout ($n = 2$) and *Lgr4/5/6* triple-knockout ($n = 2$) embryos (indicated by black arrow heads). **e**, Properly vascularized placenta of a *Lgr4/5/6* triple-knockout embryo compared to a *Lgr6*-knockout embryo. All obtained embryos (see **a**) had a similiar phenotype. Scale bar, 1 mm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



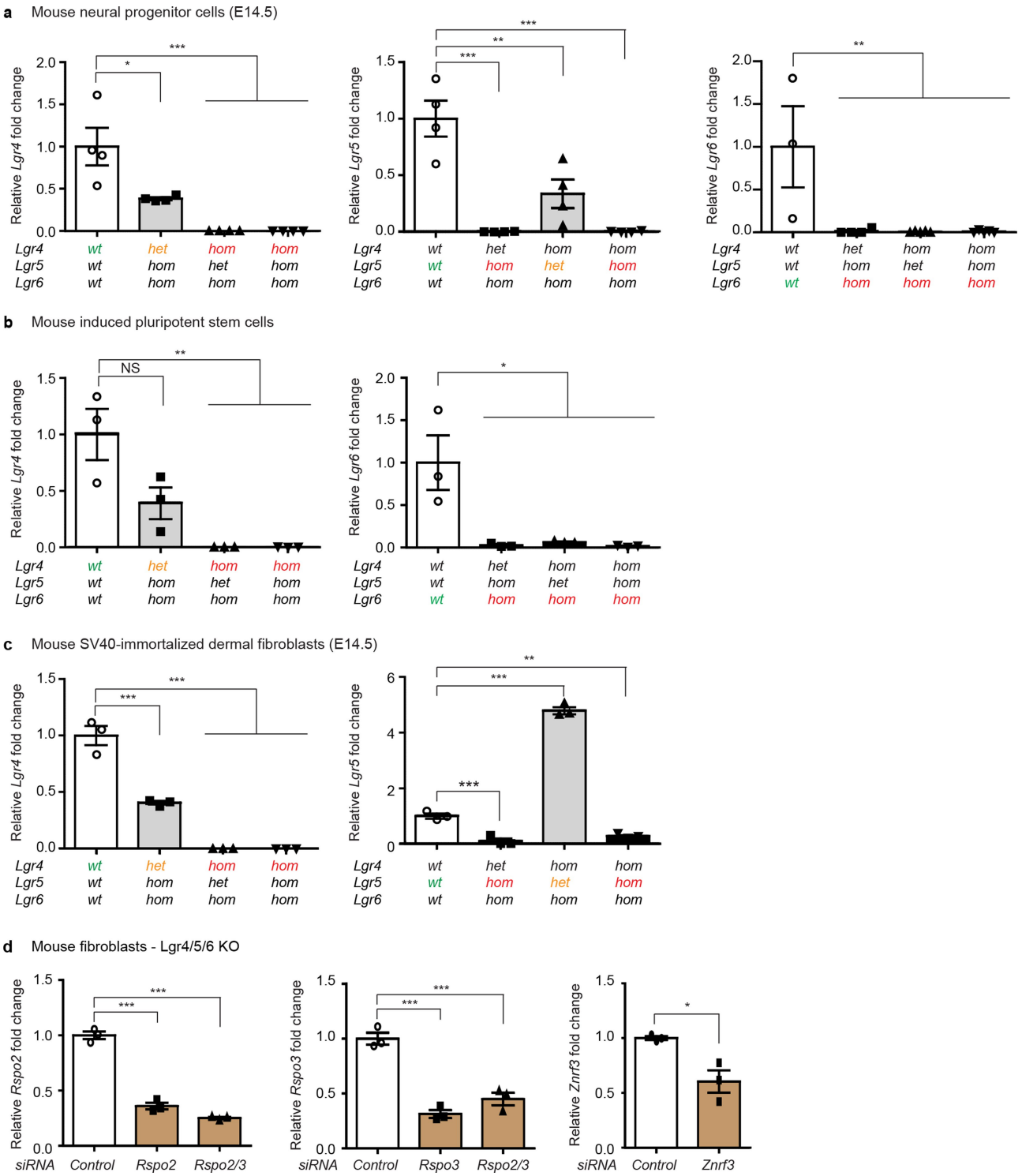
Extended Data Fig. 4 | Mouse *Lgr4*, *Lgr5* and *Lgr6* knock-in embryos recapitulate the knockout phenotypes. a, Illustration of the GFP knock-in in exon 1 of the *Lgr4, Lgr5* and *Lgr6* genes, which cause loss-of-function mutations. **b**, Liver weight in single $Lgr4^{-/-}$ ($n = 2$) and triple $Lgr4/5/6^{-/-}$ ($n = 2$) compared to wild-type ($n = 3$) E14.5 embryos. Data are mean ± s.e.m. NS, not significant. $^*P < 0.05$, one-way ANOVA test with Bonferroni's correction. **c**, $Lgr4^{-/-}$ embryos ($n = 4$) have a smaller liver. Liver weight is indicated. Scale bar, 1 mm. **d**, $Lgr4^{-/-}$ embryos ($n = 2$) show female-to-male sex reversal. Blue arrowheads point to male-specific coelomic vessels. Genetic genders (XY or XX) are indicated. Scale bar, 0.1 mm. **e**, H&E staining of coronal sections of the heads through the oral cavity. Scale bar, 1 mm. Close-ups of the tongue illustrate the ankyloglossia phenotype (tongue attached to the mouth floor, black arrow heads) in $Lgr5^{-/-}$ embryos ($n = 4$), whereas the tongue is detached for other genotypes (white arrowheads).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



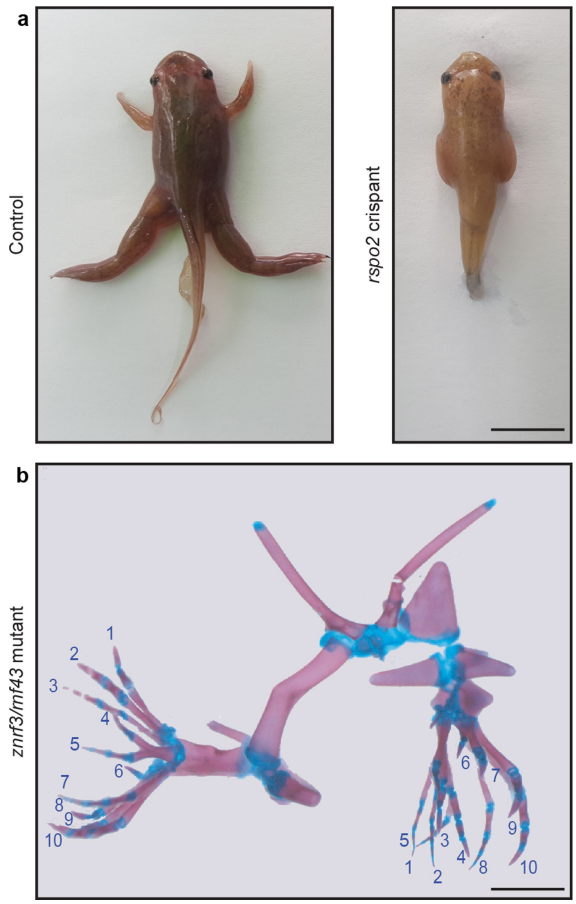
Extended Data Fig. 5 | Mouse *Lgr4*, *Lgr5* and *Lgr6* knock-in cause loss-of-function mutations. **a**, **b**, qPCR analyses for *Lgr4*, *Lgr5* and *Lgr6* expression in E14.5 limbs (**a**) and lungs (**b**) of wild-type, heterozygous and homozygous mutant embryos for the respective genes. **c**, **d**, qPCR analyses for *Lgr4* and *Lgr5* in embryonic intestine (**c**) and embryonic liver (**d**) of embryos with indicated genotype. **e**, qPCR analyses for *Lgr6* in embryonic and adult skin of wild-type versus homozygous animals. *n* indicates number of embryos. Data are mean ± s.e.m. NS, not significant. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, one-way ANOVA test with Bonferroni's correction or two-tailed unpaired *t*-test with Welch's correction when less than three groups were compared (for *Lgr6* qPCR analysis). **f**, Duplex RNAscope for the indicated gene (blue) and *Rspo2* (pink) in transverse sections of wild-type E14.5 lungs. Strongly expressed genes are denoted in bold (summary on the right). Scale bars, 0.2 mm. Experiment repeated with three different wild-type embryos.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Expression analyses in cells derived from mutant embryos. a, qPCR analyses for *Lgr4*, *Lgr5* and *Lgr6* in NPCs derived from embryos of indicated genotypes. $n = 4$ biological replicates. **b**, qPCR analyses for *Lgr4* and *Lgr6* in iPS cells derived from NPCs of indicated genotypes. $n = 3$ biological replicates. **c**, qPCR analyses for *Lgr4* and *Lgr5* in SV40-immortalized dermal fibroblasts derived from embryos of indicated genotypes. $n = 3$ biological replicates. **d**, qPCR analyses for *Rspo2*, *Rspo3* and *Znrf3* in *Lgr4/5/6* triple-knockout SV40-immortalized fibroblasts, transfected with indicated siRNAs. $n = 3$ biological replicates. Data are mean ± s.e.m. NS, not significant. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, one-way ANOVA test with Bonferroni's correction.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | Extreme phenotypes of *Xenopus tropicalis* mutant tadpoles. a, Control and *rspo2* CRISPR–Cas9-injected tadpole showing a complete tetra-amelia phenotype probably due to incomplete cleavage at the time of injection and leakage of the CRISPR–Cas9 between the two blastomeres ($n = 3$ froglets). Scale bar, 1 cm. **b**, Alizarin red and alcian blue staining of a double-mutant *rnf43/znrf3* TALEN-injected tadpole showing complete mirror-image diplopodia of both hindlimbs with 10 digits each ($n = 4$ froglets). Scale bar, 0.2 cm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Clinical characteristics of affected individuals with biallelic *RSPO2* mutations

	Humero-Femoral Hypoplasia with Radio-Tibial Ray Deficiency				Tetra-amelia syndrome						
	Family 1				Family 2		Family 3	Family 4	Family 5		
	II-3	II-5	II-6	II-7	II-2	II-3	II-1	II-1	II-1	II-2	II-3
Gender	male fetus	female fetus	female fetus	male fetus	male	male fetus	female fetus	female fetus	male fetus	n/a	Female fetus
Termination of pregnancy	30th gw	26th gw	30th gw	15th gw	n/a	26th gw	22nd gw	20th gw	20th gw	13th gw	12th gw
Year of birth	2001	2009	2011	2012	1989	1991	2006	2013	2015	2016	2017
Age at death	n/a	n/a	n/a	n/a	stillbirth	n/a	n/a	n/a	n/a	n/a	n/a
Origin	Turkey				Turkey		Portugal	France		India	
Consanguinity	yes				yes		Not reported	yes		yes	
Case report	This paper				Basaran *et al.*, 1994		Sousa *et al.*, 2008	This paper		This paper	
***RSPO2* cDNA (NM_178565)**	c.205C>T				c.208C>T		154 kb deletion chr8:108,809,266-108,963,256	c.409G>T		c.123delG	
RSPO2 protein (NP_848660)	p.Arg69Cys				p.Gln70*		n/a	p.Glu137*		p.Gly42Valfs*49	
Zygosity	parents are heterozygous	homozygous	homozygous		parents are heterozygous		homozygous	homozygous	homozygous	parents are heterozygous	homozygous
Symmetry of the limb phenotype	+	-	+	-	+	+	+	+	+	+	-
Upper limbs	Short and underdeveloped	Hypoplastic upper limb girdles. Flexion contractures of the elbow joints with pterygia. Short arms with a single tubular bone, ending with a single finger-like appendage with a well-formed nail structure.	Bilateral severely hypoplastic upper extremities with one, markedly thin tubular bone, with bilateral aplasia of thumbs and two finger-like appendages.	Hypoplastic shoulders. On the right: bowed and short humerus, single tubular bone on the forearm, thumb aplasia, ending with two finger like appendages. On the left: short arm with a single tubular bone, ending with a single finger like appendage.	Amelia	Complete absence	Complete absence	Complete absence	Complete absence	Rudimentary appendages	Complete absence of left upper limb, hypoplastic right arm with absence of forearm and hand.
Lower limbs	Short and underdeveloped	Flexion contractures of the elbow joints with pterygia. Mesomelic shortness of the legs with tibial hemimelia. Bilateral clubfeet, bilateral absence of halluces with three toes on the right foot (digit I and digit II missing).	Severely hypoplastic lower limbs with single, markedly thin tubular bone. Two toes on the feet bilaterally.	Hypoplastic pelvis, severely hypoplastic lower limbs with a single tubular bone, three toes on the left and two toes on the right, pes equinovarus.	Amelia	Rudimentary milimetric buds	Complete absence	Rudimentary appendages	Rudimentary appendages	Both femur seen, absence of lower extremities below the knees.	Absence of lower extrimities below the knee.
Others	n/a	Hypoplastic scapulae, a rudimentary triangular bone on the left, diaghragmatic hernia on antenatal ultrasound at the 27th week.	Severely hypoplastic pelvis structure. Prominent glabella, mild retrognathia.	n/a	Cleft lip-palate, no autopsy.	Bilateral cleft lip-palate, bilateral lung agenesis, bilateral palpebral fusion, micrognathia, unfused maxillary processes, severe mandibular hypoplasia, absence of nipples, small penis, absence of scrotum (testes intra-abdominal), heart defects.	Bilateral cleft lip-palate, bilateral lung agenesis, hyperecogenic focus at the right ventricle.	Bilateral cleft lip-palate, microretrognatism. Short frenulum with the tongue tethered to the floor of the mouth (ankyloglossia). Bilateral lung agenesis	Dysmorphic right ear, hypertelorism, bilateral cleft lip, complete cleft palate and severe micrognathia. Short frenulum with the tongue tethered to the floor of the mouth (ankyloglossia). Complete agenesis of both lungs and blind ending main bronchi. Agenesis of branches of pulmonary atery. Hypoplastic pulmonary veins.	Ultrasound examination: Severe micrognathia. Heart fills up most of the chest. Stomach bubble seen.	Severe mirco-retrognathia, unilateral cleft lip (left), posterior cleft palate, glossoptosis. Complete agenesis of both lungs and blind ending main bronchi. Agenesis of branches of pulmonary atery. Hypoplastic pulmonary veins.

n/a, not applicable.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | List of primers

a Human genotyping primers	
RSPO2-exon2-F	5'-CGGCTCGTGCTAGGCAGT-3'
RSPO2-exon2-R	5'-CAGGGTCCTAAAGGTGGGGA-3'
RSPO2-exon3-F	5'-GACATCCCCATGAGCCA-3'
RSPO2-exon3-R	5'-CACCCAGCAAGCTTAAACT-3'
RSPO2-exon4-F	5'-GTTGAAAGAGACAGGGATGACT-3'
RSPO2-exon4-R	5'-CAGTTCTACTGAACAAGAGAACCA-3'
RSPO2-exon5-F	5'-GGTCTTCAAGGCTGTACCACT-3'
RSPO2-exon5-R	5'-GAAGCACACAGCACACAGT-3'
RSPO2-exon6-F	5'-GGTGATGTTTTCCAGATGGGCT-3'
RSPO2-exon6-R	5'-CTGGGAACAGATACTGGGCA-3'

b Mouse genotyping primers	
Lgr4-WT-F	5'-TGCAACCCTAGAAGGGAAAA-3'
Lgr4-WT-R	5'-CTCACAGTGCTTGGGTGAAG-3'
Lgr4-null-F	5'-GCCTGCATTACCGGTCGATGCAACGA-3'
Lgr4-null-R	5'-CTCACAGTGCTTGGGTGAAG-3'
Lgr5-WT-F	5'-ACATGCTCCTGTCCTTGCT-3'
Lgr5-WT-R	5'-GTAGGAGGTGAAGACGCTGA-3'
Lgr5-null-F	5'-CACTGCATTCTAGTTGTGG-3'
Lgr5-null-R	5'-CGGTGCCCGCAGCGAG-3'
Lgr6-WT-F	5'-CGCTCGCCCGTCTGAGC-3'
Lgr6-WT-R	5'-GCGTCCAGGTCCGCAGG-3'
Lgr6-null-F	5'-CGCTCGCCCGTCTGAGC-3'
Lgr6-null-R	5'-CCTGGACGTAGCCTTCGGGC-3'

c Mouse qPCR primers	
Lgr4-qPCR-F	5'-GCCTTCACCCAAGCACTG-3'
Lgr4-qPCR-R	5'-CAGCCAGTTGTAGCTCCTCT-3'
Lgr5-qPCR-F	5'-ACAACCCCATCCAATTTGTTG-3'
Lgr5-qPCR-R	5'-CGAGGCACCATTCAAAGTCA -3'
Lgr6-qPCR-F	5'-CCCTGACTATGCCTTCCAGA-3'
Lgr6-qPCR-R	5'-ATGCTGGATGCGGTTGTTAT-3'
Rnf43-qPCR-F	5'-GGCCTATGTGTGGATTGAGC-3'
Rnf43-qPCR-R	5'-TGAGGCCAGGATGATCACAA-3'
Znrf3-qPCR-F	5'-CATCCGACTGTGCCATCTGT-3'
Znrf3-qPCR-R	5'-GCCATGGATCCACACACTTC-3'
Axin2-qPCR-F	5'-GAGTGGACTTGTGCCGACTTCA-3'
Axin2-qPCR-R	5'-GGTGGCTGGTGCAAAGACATAG-3'
Rspo1-qPCR-F	5'-ATACTTTGATGCCCGCAACC-3'
Rspo1-qPCR-R	5'-CTCACAGTGCTCGATCTTGC-3'
Rspo2-qPCR-F	5'-CGAGCCCCAGATATGAACAG-3'
Rspo2-qPCR-R	5'-AAAAGCCTACTTTGCACTTCG-3'
Rspo3-qPCR-F	5'-TGTGTCTCTCTTCGTGTCCA-3'
Rspo3-qPCR-R	5'-AGGTATCACAGTCAACTTTGCA-3'
Rspo4-qPCR-F	5'-GGACATGCTCGCCCTGTA-3'
Rspo4-qPCR-R	5'-GAACAGCCATTCTCCTCCGA-3'
Actin-qPCR-F	5'-AAGGCCAACCGTGAAAAGAT-3'
Actin-qPCR-R	5'-GTGGTACGACCAGAGGCATAC-3'

d Target sequences for *Xenopus tropicalis* rnf43 and znrf3 TALENs, and rspo2 gRNA	
Xt_rnf43-TALEN-ELD	5'-TGCTCACGGTGACTCTC-3'
Xt_rnf43-TALEN-KKR	5'-CCATGGGCACCACGGAA-3'
Xt_znrf3-TALEN-ELD	5'-TTTTTCGTGGTGGTGTC-3'
Xt_znrf3-TALEN-KKR	5'-CTCCTTATCAAGATCAA-3'
Xt_rspo2-gRNA	5'-TGACTCCATAGTATCCAGGAGGG-3'

e PCR primers for *Xenopus tropicalis in situ* probes	
Xt_znrf3-insitu-F	5'-AATTAACCCTCACTAAAGGG**GCTGTGATATTTGATGTGTCTG-3'**
Xt_znrf3-insitu-R	5'-TAATACGACTCACTATAGGG**ACTTCCACCAACCTCCTG-3'**
Xt_rnf43-insitu-F	5'-AATTAACCCTCACTAAAGGG**GGCTTCATTTCCATTGTCAAACTG-3'**
Xt_rnf43-insitu-R	5'-TAATACGACTCACTATAGGG**TCCTGCCCATCTGTGAACTC-3'**

f *Xenopus tropicalis* genotyping primers	
Xt_rnf43-F	5'-CCACACCCCAACAAAATCA-3'
Xt_rnf43-R	5'-CCACACCCCAACAAAATCA-3'
Xt_znrf3-F	5'-ACAGCATGCCTTCCCTACAC-3'
Xt_znrf3-R	5'-GTAGGTTGCTGCCAAATCTCAC-3'
Xt_rspo2-F	5'-GTCGTGTTGAAATGGTGCGG-3'
Xt_rspo2-R	5'-GTTCCTTGACAAGTATCCAAGCTG-3'

a, Genotyping primers for human *RSPO2*. **b**, Genotyping primers for mouse *Lgr4, Lgr5* and *Lgr6* wild-type and knock-in alleles. **c**, Mouse qPCR primers used in this study. **d**, Target sequences for *X. tropicalis rnf43* and *znrf3* TALENs, and *rspo2* gRNA. **e**, PCR primers for *X. tropicalis* in situ probes via cloning-free methodology. **f**, Genotyping primers for *X. tropicalis rnf43* and *znrf3* TALENs, and *rspo2* CRISPR–Cas9 target sites.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0121-3

Mxra8 is a receptor for multiple arthritogenic alphaviruses

Rong Zhang[1], Arthur S. Kim[1,2], Julie M. Fox[1], Sharmila Nair[1], Katherine Basore[3], William B. Klimstra[4], Rebecca Rimkunas[5], Rachel H. Fong[5], Hueylie Lin[1], Subhajit Poddar[1], James E. Crowe Jr[6], Benjamin J. Doranz[5], Daved H. Fremont[3] & Michael S. Diamond[1,2,3,7]*

Arthritogenic alphaviruses comprise a group of enveloped RNA viruses that are transmitted to humans by mosquitoes and cause debilitating acute and chronic musculoskeletal disease[1]. The host factors required for alphavirus entry remain poorly characterized[2]. Here we use a genome-wide CRISPR–Cas9-based screen to identify the cell adhesion molecule Mxra8 as an entry mediator for multiple emerging arthritogenic alphaviruses, including chikungunya, Ross River, Mayaro and O'nyong nyong viruses. Gene editing of mouse *Mxra8* or human *MXRA8* resulted in reduced levels of viral infection of cells and, reciprocally, ectopic expression of these genes resulted in increased infection. Mxra8 bound directly to chikungunya virus particles and enhanced virus attachment and internalization into cells. Consistent with these findings, Mxra8–Fc fusion protein or anti-Mxra8 monoclonal antibodies blocked chikungunya virus infection in multiple cell types, including primary human synovial fibroblasts, osteoblasts, chondrocytes and skeletal muscle cells. Mutagenesis experiments suggest that Mxra8 binds to a surface-exposed region across the A and B domains of chikungunya virus E2 protein, which are a speculated site of attachment. Finally, administration of the Mxra8–Fc protein or anti-Mxra8 blocking antibodies to mice reduced chikungunya and O'nyong nyong virus infection as well as associated foot swelling. Pharmacological targeting of Mxra8 could form a strategy for mitigating infection and disease by multiple arthritogenic alphaviruses.

We performed a genome-wide screen for host factors required for chikungunya virus (CHIKV) infection, using the CRISPR–Cas9 system[3,4] and lentiviruses delivering single-guide RNAs (sgRNA) targeting 20,611 mouse genes (Extended Data Fig. 1a). We inoculated lentivirus-transduced 3T3 mouse fibroblasts with CHIKV strain 181/25 that contained the mKate2 reporter (CHIKV-181/25-mKate2), such that almost all cells expressed the reporter gene by 24 h. The few cells that lacked mKate2 expression were sorted, propagated in the presence of neutralizing anti-CHIKV monoclonal antibodies (mAbs)[5] and then re-inoculated with CHIKV-181/25-mKate2. After two rounds of infection and sorting, genomic DNA from mKate2-negative cells was collected, sgRNAs were sequenced and then analysed using MAGeCK[6] (Supplementary Tables 1, 2). The top candidate was *Mxra8* (also known as *DICAM*, *ASP3* or limitrin), an adhesion molecule found in mammals, birds and amphibians (Extended Data Fig. 1b, c) that is expressed on epithelial, myeloid and mesenchymal cells[7–10] and shares homology with junctional adhesion molecule[9], a reovirus entry receptor[11]. We validated *Mxra8* using three different sgRNAs in bulk 3T3 cells, by generating Δ*Mxra8* single-cell clones in 3T3 and MEF cells, and by confirming gene deletion and cell viability (Extended Data Fig. 2a–e). Infection of CHIKV-181/25 was reduced in Δ*Mxra8* cells, and trans-complementation of *Mxra8* in Δ*Mxra8* 3T3 cells restored infectivity (Fig. 1a, b). As CHIKV-181/25 is a cell-culture-adapted vaccine strain[12] that has acquired heparan sulfate binding activity[13], we evaluated Mxra8 with other CHIKV strains. Infection of CHIKV-AF15561—the parental Asian strain of CHIKV-181/25, which binds poorly to heparan sulfate[14]—and CHIKV-37997, a West African genotype strain, was abolished in Δ*Mxra8* 3T3 cells, reduced in Δ*Mxra8* MEF cells (Fig. 1a) and restored in trans-complemented Δ*Mxra8* 3T3 cells (Fig. 1b, c). However, the dependence on Mxra8 was less with CHIKV-LR 2006, a strain of the East/Central/South African genotype (Fig. 1a, d). To confirm that CHIKV required Mxra8 independently of heparan sulfate binding, we expressed murine Mxra8 in parental or glycosaminoglycan-deficient Chinese hamster ovary cells[15] (Extended Data Fig. 3a). Expression of Mxra8 enhanced infectivity of CHIKV regardless of whether Chinese hamster ovary cells expressed heparan sulfate or other glycosaminoglycans (Extended Data Fig. 3b, c).

We tested the requirement of Mxra8 for infection by other alphaviruses. Whereas Mayaro, Ross River, O'nyong nyong and Barmah Forest arthritogenic alphaviruses showed reduced infection in Δ*Mxra8* 3T3 cells, Semliki Forest and Getah viruses had partial phenotypes, and other related alphaviruses (Sindbis, Bebaru, Una and Middleburg viruses) showed little dependence on Mxra8 (Fig. 1e and Extended Data Fig. 4). Minimal differences in infection were observed between control and Δ*Mxra8* 3T3 cells with chimaeric Sindbis viruses expressing the structural genes of the Eastern or Western equine encephalitis alphaviruses, or a Venezuelan equine encephalitis virus that expressed GFP (Fig. 1e). No effect of Mxra8 was seen on infection of unrelated positive- or negative-sense RNA viruses (Fig. 1f). We next assessed whether any of the four isoforms of the human *MXRA8* orthologue (Extended Data Fig. 1b) served a similar function. As HeLa cells did not express MXRA8 (Extended Data Fig. 5), we used these cells for ectopic expression (Extended Data Fig. 6a). MXRA8-1, MXRA8-2 and MXRA8-4—but not MXRA8-3—were detected on the cell surface, and MXRA8-1 and MXRA8-2 enhanced CHIKV infectivity (Fig. 1g). Similarly, expression of MXRA8-2 in A549 or 293T cells resulted in greater CHIKV infection (Extended Data Fig. 6b, c). Consistent with this observation, expression of different sgRNAs targeting all isoforms of *MXRA8* in MRC-5 human lung fibroblasts, HFF-1 foreskin fibroblasts, RPE retinal pigment epithelial cells and Hs 633T fibrosarcoma cells resulted in less CHIKV infection than in control gene-edited cells (Fig. 1h and Extended Data Fig. 7a, b).

To determine whether Mxra8 is required for replication, we transfected CHIKV genomic RNA into control and Δ*Mxra8* MEF cells in the presence of NH_4Cl to inhibit virus maturation and further rounds of infection. As no difference in CHIKV gene expression was detected (Fig. 2a), Mxra8 does not appear to affect translation or replication. We demonstrated an Mxra8-dependence of the structural proteins using pseudotyped viruses. Whereas infection of CHIKV pseudotyped virions that encapsidated a murine leukaemia virus GFP-reporter RNA

[1]Department of Medicine, Washington University School of Medicine, Saint Louis, MO, USA. [2]Department of Molecular Microbiology, Washington University School of Medicine, Saint Louis, MO, USA. [3]Department of Pathology and Immunology, Washington University School of Medicine, Saint Louis, MO, USA. [4]Center for Vaccine Research, University of Pittsburgh, Pittsburgh, PA, USA. [5]Integral Molecular Inc., Philadelphia, PA, USA. [6]Vanderbilt Vaccine Center, Department of Pediatrics and Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, USA. [7]Andrew M. and Jane M. Bursky Center for Human Immunology and Immunotherapy Programs, Washington University School of Medicine, Saint Louis, MO, USA.
*e-mail: diamond@wusm.wustl.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

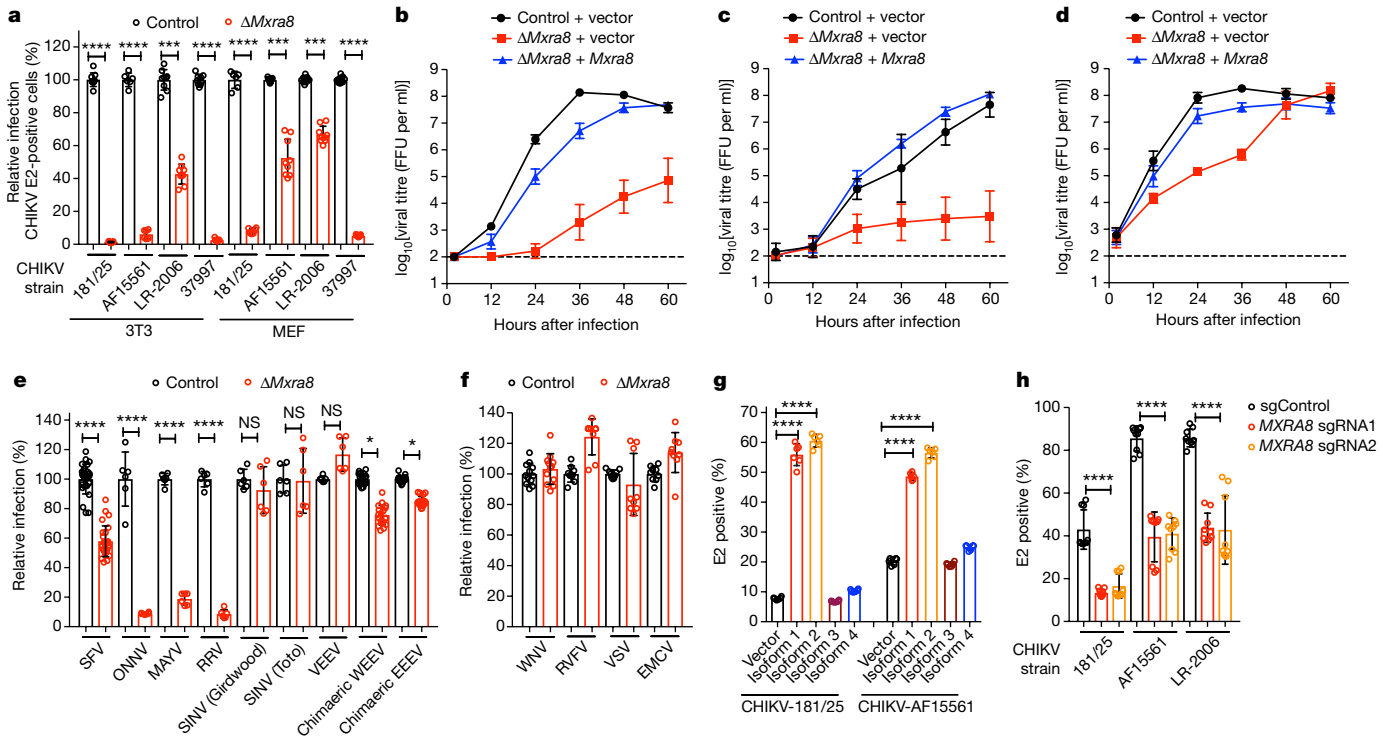


Fig. 1 | Mxra8 is required for optimal infection of CHIKV and other alphaviruses. a, Δ*Mxra8* or control 3T3 or MEF cells were inoculated with CHIKV and stained for E2 protein (3 experiments, $n = 9$; two-tailed *t*-test with Holm–Sidak correction; mean ± s.d.). **b**–**d**, Multi-step growth curves with CHIKV-181/25 (**b**), CHIKV-AF15561 (**c**) or CHIKV-LR-2006 (**d**) in control, Δ*Mxra8* or *Mxra8* trans-complemented Δ*Mxra8* 3T3 cells (3 experiments, $n = 9$; mean ± s.d.). FFU, focus forming units. **e**, Δ*Mxra8* or control 3T3 cells were inoculated with alphaviruses and processed for E2 or reporter gene expression (3 or more experiments, $n = 6$ except for Semliki Forest virus (SFV), Sindbis–Western equine encephalitis virus chimaera (WEEV) and Sindbis–Eastern equine encephalitis virus chimaera (EEEV) in which $n = 18$; two-tailed *t*-test with Holm–Sidak correction; mean ± s.d.). **f**, Δ*Mxra8* or control 3T3 cells were inoculated with indicated viruses and processed for viral antigen or reporter gene expression (3 experiments, mean ± s.d.). **g**, HeLa cells were transduced with control or *MXRA8-1*, *MXRA8-2*, *MXRA8-3* or *MXRA8-4* alleles, inoculated with CHIKV and processed for E2 staining (3 experiments, $n = 6$; one-way ANOVA with Dunnett's test; mean ± s.d.). **h**, Human MRC-5 cells depleted of *MXRA8* with two different sgRNA were inoculated with CHIKV, and E2 expression was analysed (3 experiments, $n = 9$; one-way ANOVA with Dunnett's test; mean ± s.d.). EMCV, encephalomyocarditis virus; MAYV; Mayaro virus; ONNV, O'nyong nyong virus; RRV, Ross River virus; RVFV, Rift Valley fever virus; VEEV, Venezuelan equine encephalitis virus; VSV, vesicular stomatitis virus; WNV, West Nile virus. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$; NS, not significant.

was reduced in Δ*Mxra8* compared to control MEF cells, infection of Eastern or Western equine encephalitis virus pseudotyped virions was not (Fig. 2b). Consistent with a role for Mxra8 in the entry pathway, at 4 °C CHIKV-AF15561 showed reduced binding to Δ*Mxra8* compared to control MEF cells, and increased binding to cells overexpressing *Mxra8* (Fig. 2c, d). When virus internalization assays were performed at 37 °C, less CHIKV RNA was measured within Δ*Mxra8* MEF cells, and more CHIKV RNA was detected in cells overexpressing *Mxra8* (Fig. 2c).

To corroborate an effect of Mxra8 on binding and entry, we generated Fc fusion proteins with the extracellular domains of mouse Mxra8 (Mxra8–Fc) or human MXRA8-2 (MXRA8-2–Fc), along with a control osteoprotegerin protein (OPG–Fc) (Extended Data Fig. 8a). Pre-incubation with Mxra8–Fc or MXRA8-2–Fc, but not the control OPG–Fc, reduced CHIKV-181/25 infection in 3T3 (Fig. 2e) and MRC-5 (Fig. 2f) cells. We tested a panel of hamster mAbs against mouse Mxra8 (Extended Data Fig. 8b) for their capacity to inhibit CHIKV infection. Seven mAbs bound to mouse Mxra8–Fc, with four also recognizing human MXRA8-2–Fc (Extended Data Fig. 8c). Pre-treatment of 3T3 cells with anti-Mxra8 mAbs reduced CHIKV infection (Fig. 2g). Three of the mAbs that bound human MXRA8-2–Fc also reduced infection in MRC-5 cells (Extended Data Fig. 8d). To establish the importance of the Mxra8 ectodomain, we engineered forms with glycophosphatidylinositol anchors (Mxra8-GPI) or lacking a cytoplasmic domain (Mxra8-ΔC-tail) for trans-complementation (Extended Data Fig. 9). Notably, Mxra8-GPI and Mxra8-ΔC-tail restored CHIKV infection (Fig. 2h). Interaction with the Mxra8 ectodomain may facilitate viral glycoprotein conformational changes that are required for internalization or fusion[16] or potentiate interactions with other host factors that bridge membrane penetration and entry[17]. Finally, heterologous expression of human *MXRA8-2* in Δ*Mxra8* mouse cells also restored CHIKV infection (Fig. 2h).

To determine whether Mxra8 directly binds to CHIKV, we captured virions or virus-like particles[18] with a human anti-CHIKV mAb[19], and added Mxra8–Fc or MXRA8-2–Fc in an enzyme-linked immunosorbent assay. Both Mxra8–Fc and MXRA8-2–Fc, but not OPG–Fc, bound to CHIKV virions and virus-like particles (Fig. 3a, b). By comparison, Mxra8–Fc and MXRA8-2–Fc did not bind efficiently to Eastern equine encephalitis virus particles derived from a chimaera with Sindbis virus (Fig. 3c). In a complementary assay, we assessed binding of Mxra8–Fc protein to cell-surface-displayed alphavirus proteins on infected cells[19,20]. Mxra8–Fc bound cells infected with CHIKV, O'nyong nyong virus, Mayaro virus and Ross River virus, but not cells infected with Sindbis virus or Venezuelan equine encephalitis virus (Extended Data Fig. 10a, b). We also analysed binding of Mxra8 to CHIKV virus-like particles by surface plasmon resonance, and found a slow association rate, a long half-life and an affinity of about 200 nM (Fig. 3d).

We evaluated whether human anti-CHIKV mAbs that bound epitopes within the E2 protein[19] altered Mxra8–Fc binding to CHIKV. Several mAbs recognizing epitopes in the A domain (2H1, 8G18, 3E23 and 1O6) and mAbs recognizing shared epitopes in the A and B domains (1H12 and 4J14) inhibited binding, whereas other mAbs that localize to distinct sites had less effect (Fig. 3e). We then tested

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

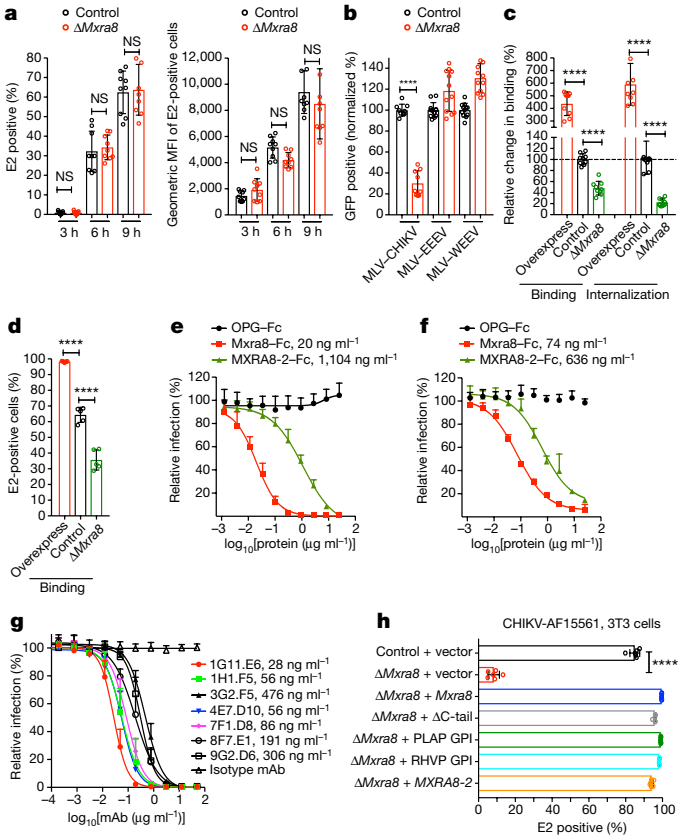


Fig. 2 | Mxra8 modulates CHIKV attachment and internalization. **a**, Transfection of CHIKV RNA into control or Δ*Mxra8* cells. Cells were analysed for E2 expression: left, percent positive; right, mean fluorescence intensity (MFI) (3 experiments, $n = 9$; two-tailed t-test with Holm–Sidak correction, mean ± s.d.). **b**. Murine leukaemia virus (MLV) RNAs encoding GFP and pseudotyped with alphavirus structural genes were added to control or Δ*Mxra8* cells. Data are from three (MLV-CHIKV, $n = 9$) or four (MLV-WEEV and MLV-EEEV, $n = 12$) experiments (two-tailed t-test with Holm–Sidak correction; mean ± s.d.). **c**, **d**, CHIKV-AF15561 was incubated with control, Δ*Mxra8* and *Mxra8*-overexpressing MEF cells at 4 °C (**c** left, **d**) or 37 °C (**c** right) as described in Methods. Cells were collected, and RNA (CHIKV and *Gapdh*) was measured by qRT–PCR (**c**) or surface E2 protein was analysed by flow cytometry (**d**) (3 experiments; **c**, $n = 9$; **d**, $n = 5$; one-way ANOVA with Dunnett's test; mean ± s.d.). **e**, **f**, Mxra8–Fc, MXRA8-2–Fc or OPG–Fc were mixed with CHIKV-181/25 before infection of 3T3 (**e**) or MRC-5 (**f**) cells. Data are from two (MRC-5, $n = 6$) or three (3T3, $n = 8$–12) experiments; mean ± s.d.; 50% inhibitory concentrations (EC50) are indicated to the right of each protein. **g**, Blockade of CHIKV-181/25 infection in 3T3 cells with hamster anti-Mxra8 or isotype control mAbs (2 experiments, $n = 6$; mean ± s.d., EC50 values are indicated to the right of each mAb). **h**, Trans-complementation of Δ*Mxra8* cells with vector, *Mxra8*, *Mxra8* ΔC-tail, *Mxra8* with GPI anchors (derived from placental alkaline phosphatase (PLAP) or the rodent herpesvirus Peru (RHVP) open reading frame R1 gene[26]) or human *MXRA8-2* (3 experiments; $n = 6$; one-way ANOVA with Dunnett's test; mean ± s.d.). ****$P < 0.0001$; NS, not significant.

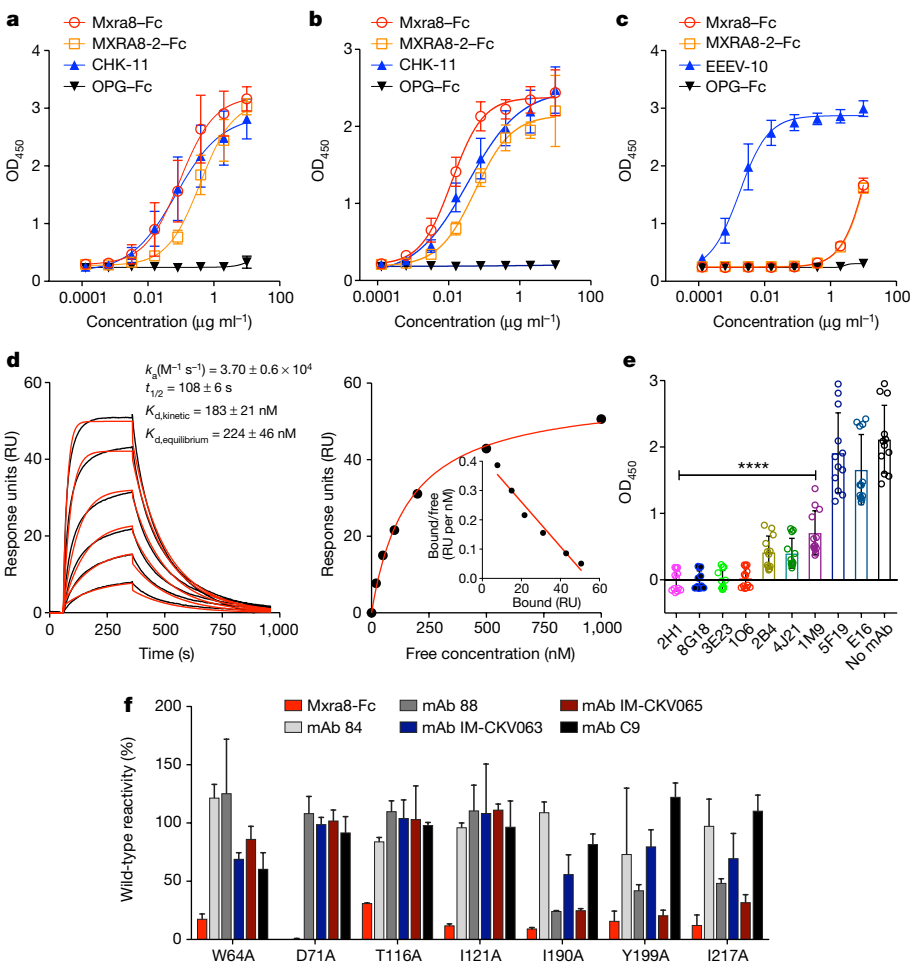


Fig. 3 | Direct binding of Mxra8 to CHIKV. **a**–**c**, Purified CHIKV-181/25 (**a**), CHIKV virus-like particles (**b**) or chimaeric EEEV virions (**c**) were captured with anti-CHIKV or anti-EEEV human mAbs. Mxra8–Fc, MXRA8-2–Fc or OPG–Fc, and positive-control mAbs (CHK-11 or EEEV-10), were added. Data are from two (**b**, **c**) or three (**a**) experiments (**a**, $n = 8$; **b**, $n = 4$; **c**, $n = 6$); mean ± s.d. **d**, Left, sensograms of Mxra8 binding to CHIKV virus-like particles. Experimental curves (black traces) were fit using a 1:1 binding model (red traces). Right, representative response curve for steady-state analysis, in which binding is plotted versus Mxra8 concentration. Inset, linear Scatchard plot (4 experiments; mean and s.e.m.). **e**. Antibody blockade of Mxra8–Fc binding to CHIKV. Virus was incubated with indicated human mAbs against CHIKV, a mAb against WNV (E16) or no mAb before the addition of Mxra8–Fc (4 experiments, $n = 12$; one-way ANOVA with Dunnett's test; mean ± s.d.). **f**, Residues that result in loss of Mxra8–Fc binding to cell-surface-displayed CHIKV E2–E1. Residues are considered involved in the epitope if there is diminished binding without loss of protein integrity as judged by retention of interaction with mAbs (except for those previously mapped to a specific residue) (see also Supplementary Table 3). Data are from two experiments. ****$P < 0.0001$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

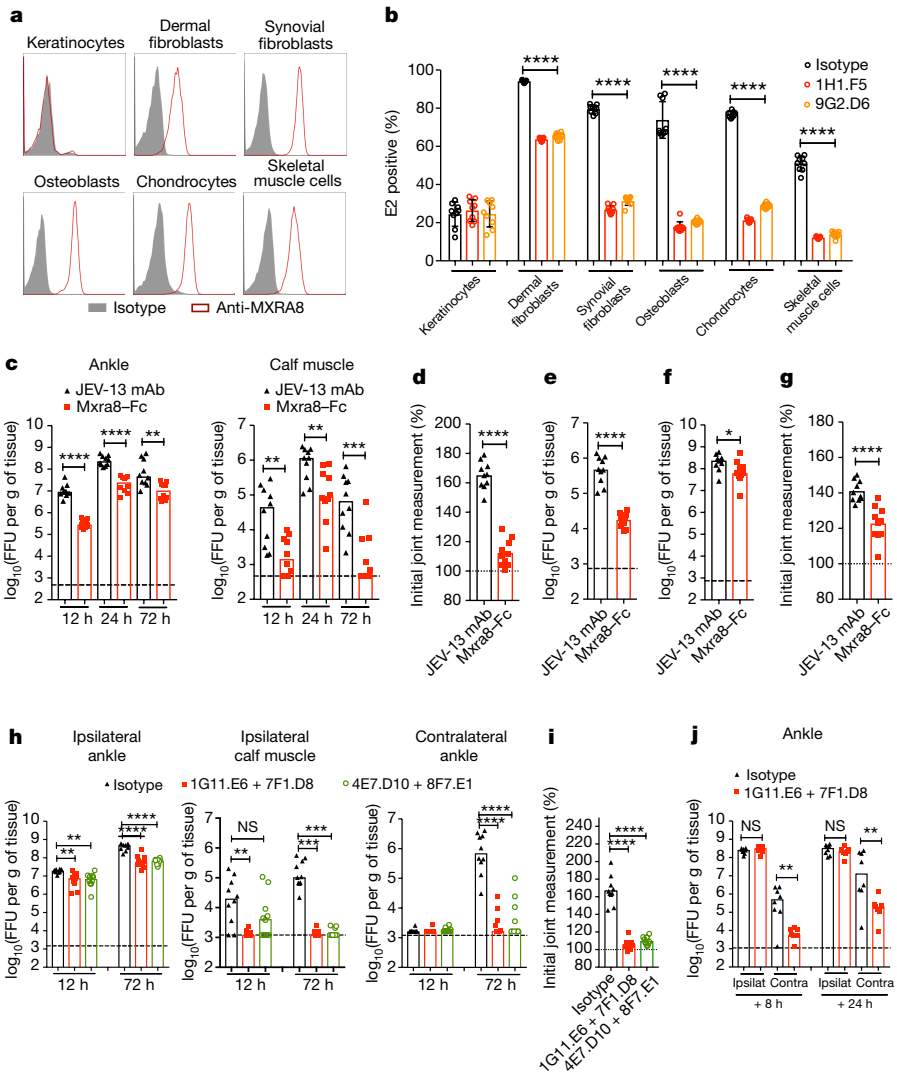


Fig. 4 | Mxra8 contributes to alphavirus pathogenesis. a, Surface expression of MXRA8 on primary human keratinocytes, dermal fibroblasts, synovial fibroblasts, osteoblasts, chondrocytes and skeletal muscle cells. One experiment of three is shown. **b**, Primary human cells were pre-incubated with anti-MXRA8 blocking mAbs before addition of CHIKV-AF15561 and processed for E2 staining (3 experiments, $n = 9$; one-way ANOVA with Dunnett's test). **c**, **d**, Mxra8–Fc or JEV-13 mAb were incubated with CHIKV-AF15561 for 30 min before subcutaneous inoculation. **c**, At 12, 24 and 72 h, CHIKV was measured in the ipsilateral ankle and calf muscle. **d**, At 72 h, foot swelling was measured (2 experiments, $n = 10$; median viral titres: two-tailed Mann–Whitney test; mean foot swelling: two-tailed unpaired *t*-test). **e**, Mxra8–Fc or JEV-13 mAb was mixed with O'nyong nyong virus immediately before subcutaneous inoculation. At 12 h, O'nyong nyong virus was measured in the ipsilateral ankle (2 experiments, $n = 10$; two-tailed unpaired *t*-test; median values). **f**, **g**, Mxra8–Fc or JEV-13 mAb was administered via an intraperitoneal route 6 h before CHIKV-AF15561 inoculation in the footpad. At 24 h, CHIKV was measured in the ipsilateral ankle (**f**). At 72 h, foot swelling was measured (**g**) (2 experiments, $n = 10$; median viral titres: two-tailed Mann–Whitney test; mean foot swelling: two-tailed unpaired *t*-test). **h**–**j**, Pairs of anti-Mxra8 mAbs or isotype control hamster mAbs were administered via an intraperitoneal route 12 h before (**h**, **i**) or 8 or 24 h after (**j**) inoculation of CHIKV-AF15561. At 12 (**h**) and 72 (**h**, **j**) h, CHIKV was measured. At 72 h, foot swelling (**i**) was measured (2 experiments; **h** left, **i**: $n = 10$; one-way ANOVA with Dunnett's test; **h** middle and right: $n = 10$; Kruskal–Wallis with Dunn's test; **j**: $n = 8$; two-tailed Mann–Whitney test). Ipsilat, ipsilateral; contra, contralateral. $^*P < 0.05$; $^{**}P < 0.01$; ***, $P < 0.001$; $^{****}P < 0.0001$.

the binding of Mxra8–Fc to an alanine scanning mutagenesis library of E2 A and B domains in the context of display on the surface of 293T cells[19,20]. Residues W64, D71, T116 and I121 in the A domain, and I190, Y199 and I217 in the B domain, of E2 emerged as essential for optimal Mxra8–Fc binding (Fig. 3f and Supplementary Table 3) and overlap the binding sites of the blocking mAbs that we tested[19]. Mapping of these residues onto the p62 (E2 precursor)–E1 heterodimer or the trimer of heterodimers (Extended Data Fig. 10c, d) on the virion surface[21] revealed a solvent-accessible epitope across the top of the A and B domains, which is a proposed site of alphavirus receptor engagement[22,23].

To begin to assess the physiological importance of MXRA8 interaction with CHIKV, we tested surface expression of MXRA8 on primary human keratinocytes, dermal fibroblasts, synovial fibroblasts, osteoblasts, chondrocytes and skeletal muscle cells (Fig. 4a), all of which are targets of infection by alphaviruses[24]. Pretreatment with anti-MXRA8 blocking mAbs inhibited infection of CHIKV-AF15561 in all cells but keratinocytes, which lack MXRA8 expression (Fig. 4a, b). We next evaluated whether co-injection of Mxra8–Fc with CHIKV-AF15561 would diminish infection in C57BL/6 mice. The addition of Mxra8–Fc diminished CHIKV infection in the ipsilateral ankle and muscle (Fig. 4c) and reduced foot swelling (Fig. 4d) compared to a control mAb. Treatment with Mxra8–Fc also inhibited O'nyong nyong virus infection in the ipsilateral ankle of mice (Fig. 4e). We next administered Mxra8–Fc via an intraperitoneal route, and 6 h later we inoculated CHIKV in the footpad. Mxra8–Fc treatment reduced foot swelling and viral burden in the ipsilateral ankle (Fig. 4f, g), although the phenotype was less pronounced than

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

in co-injection experiments. To extend these findings, we transferred hamster anti-Mxra8 blocking or control mAbs to mice via an intraperitoneal route 12 h before CHIKV inoculation. Reduced CHIKV titres were observed in the ipsilateral ankle and calf muscle, and contralateral ankle at 12 and 72 h after infection in anti-Mxra8 compared to control mAb-treated mice (Fig. 4h). Treatment with anti-Mxra8 mAbs also reduced foot swelling (Fig. 4i). In post-exposure therapeutic experiments, we observed reduced CHIKV infection in the contralateral ankle and muscle when anti-Mxra8 mAbs were administered at 8 or 24 h after virus inoculation (Fig. 4j and Extended Data Fig. 8e). These in vivo experiments establish a function for Mxra8 in the pathogenesis of infection of arthritogenic alphaviruses.

Our studies establish that mouse Mxra8 contributes to CHIKV entry and is required for infection and disease. Human MXRA8 also bound CHIKV and supported infection, and MXRA8 expression in primary human cells overlapped with the tropism of CHIKV in vivo. Infection of several arthritogenic alphaviruses—including CHIKV, O'nyong nyong virus, Mayaro virus and Ross River virus—was reduced in Δ*Mxra8* cells, which suggests that Mxra8 may serve as a shared receptor. Our data contrast with those relating to natural resistance-associated macrophage protein (NRAMP2), which is an entry receptor for Sindbis virus but not for CHIKV or Ross River virus[25]. Nonetheless, residual CHIKV infection in the absence of Mxra8 in cells and in mice, and the absence of an apparent mosquito orthologue, suggests that additional unidentified host factors contribute to cell binding and entry.

Our mutagenesis mapping studies suggest that amino acids in the E2 A and B domains contribute to the interaction of CHIKV with Mxra8. Higher-resolution structural experiments are needed to define the complete footprint of binding between Mxra8 and CHIKV E2 protein. Such studies could facilitate the development of small molecules or biological agents that disrupt Mxra8 interaction with E2 protein, which could form the basis of therapeutic strategies for the amelioration of diseases caused by multiple emerging alphaviruses.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0121-3.

Received: 8 December 2017; Accepted: 3 April 2018;
Published online: 16 May 2018

1. Weaver, S. C., Charlier, C., Vasilakis, N. & Lecuit, M. Zika, Chikungunya, and other emerging vector-borne viral diseases. *Annu. Rev. Med.* **69**, 395–408 (2018).
2. Vancini, R., Hernandez, R. & Brown, D. Alphavirus entry into host cells. *Prog. Mol. Biol. Transl. Sci.* **129**, 33–62 (2015).
3. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
4. Jinek, M. et al. RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
5. Pal, P. et al. Development of a highly protective combination monoclonal antibody therapy against Chikungunya virus. *PLoS Pathog.* **9**, e1003312 (2013).
6. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).
7. Han, S. W. et al. DICAM inhibits angiogenesis via suppression of AKT and p38 MAP kinase signalling. *Cardiovasc. Res.* **98**, 73–82 (2013).
8. Jung, Y. K. et al. DICAM inhibits osteoclast differentiation through attenuation of the integrin αVβ3 pathway. *J. Bone Miner. Res.* **27**, 2024–2034 (2012).
9. Jung, Y. K. et al. DICAM, a novel dual immunoglobulin domain containing cell adhesion molecule interacts with αvβ3 integrin. *J. Cell. Physiol.* **216**, 603–614 (2008).
10. Yonezawa, T. et al. Limitrin, a novel immunoglobulin superfamily protein localized to glia limitans formed by astrocyte endfeet. *Glia* **44**, 190–204 (2003).
11. Barton, E. S. et al. Junction adhesion molecule is a receptor for reovirus. *Cell* **104**, 441–451 (2001).
12. Levitt, N. H. et al. Development of an attenuated strain of chikungunya virus for use in vaccine production. *Vaccine* **4**, 157–162 (1986).
13. Gardner, C. L., Burke, C. W., Higgs, S. T., Klimstra, W. B. & Ryman, K. D. Interferon-alpha/beta deficiency greatly exacerbates arthritogenic disease in mice infected with wild-type chikungunya virus but not with the cell culture-adapted live-attenuated 181/25 vaccine candidate. *Virology* **425**, 103–112 (2012).
14. Ashbrook, A. W. et al. Residue 82 of the Chikungunya virus E2 attachment protein modulates viral dissemination and arthritis in mice. *J. Virol.* **88**, 12180–12192 (2014).
15. Esko, J. D., Stewart, T. E. & Taylor, W. H. Animal cell mutants defective in glycosaminoglycan biosynthesis. *Proc. Natl Acad. Sci. USA* **82**, 3197–3201 (1985).
16. Lee, J. et al. Structure of the Ebola virus envelope protein MPER/TM domain and its interaction with the fusion loop explains their fusion activity. *Proc. Natl Acad. Sci. USA* **114**, E7987–E7996 (2017).
17. Doranz, B. J., Berson, J. F., Rucker, J. & Doms, R. W. Chemokine receptors as fusion cofactors for human immunodeficiency virus type 1 (HIV-1). *Immunol. Res.* **16**, 15–28 (1997).
18. Akahata, W. et al. A virus-like particle vaccine for epidemic Chikungunya virus protects nonhuman primates against infection. *Nat. Med.* **16**, 334–338 (2010).
19. Smith, S. A. et al. Isolation and characterization of broad and ultrapotent human monoclonal antibodies with therapeutic activity against Chikungunya virus. *Cell Host Microbe* **18**, 86–95 (2015).
20. Fox, J. M. et al. Broadly neutralizing alphavirus antibodies bind an epitope on E2 and inhibit entry and egress. *Cell* **163**, 1095–1107 (2015).
21. Sun, S. et al. Structural analyses at pseudo atomic resolution of Chikungunya virus and antibodies show mechanisms of neutralization. *eLife* **2**, e00435 (2013).
22. Li, L., Jose, J., Xiang, Y., Kuhn, R. J. & Rossmann, M. G. Structural changes of envelope proteins during alphavirus fusion. *Nature* **468**, 705–708 (2010).
23. Voss, J. E. et al. Glycoprotein organization of Chikungunya virus particles revealed by X-ray crystallography. *Nature* **468**, 709–712 (2010).
24. Couderc, T. & Lecuit, M. Chikungunya virus pathogenesis: From bedside to bench. *Antiviral Res.* **121**, 120–131 (2015).
25. Rose, P. P. et al. Natural resistance-associated macrophage protein is a cellular receptor for sindbis virus in both insect and mammalian hosts. *Cell Host Microbe* **10**, 97–104 (2011).
26. Loh, J. et al. Identification and sequencing of a novel rodent gammaherpesvirus that establishes acute and latent infection in laboratory mice. *J. Virol.* **85**, 2642–2656 (2011).

Acknowledgements This study was supported by NIH grants R01AI114816, R01AI123348, and R01AI095436, contract HHSN272201400058C, and the Defense Reduction Threat Agency HDTRA1-15-1-0013. We thank M. Heise, T. Morrison, M. Farzan, J. Miner and T. Dermody for discussions, reagents and comments on the manuscript.

Reviewer information *Nature* thanks S. Mahalingam, C. Richardson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.Z. performed CRISPR–Cas9 screening and infection studies with J.M.F., S.N., H.L. and S.P. A.S.K. generated Mxra8–Fc and performed ELISAs. K.B. performed SPR experiments. R.R. and R.H.F. performed mapping experiments. R.Z., A.S.K., J.M.F., B.J.D., D.H.F. and M.S.D. designed experiments. W.B.K. and J.E.C. provided key reagents. R.Z., A.S.K., J.M.F., K.B., R.H.F. and B.J.D. performed data analysis. R.Z. and M.S.D. wrote the initial manuscript draft.

Competing interests M.S.D. is a consultant for Inbios and Sanofi-Pasteur and on the Scientific Advisory Board of Moderna. J.E.C. is on the Scientific Advisory Boards of GigaGen, Meissa Vaccines, and PaxVax, and is a Founder of IDBiologics. B.J.D. is a shareholder of Integral Molecular.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0121-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0121-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.S.D.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Cells and viruses. Vero, NIH-3T3, MEF, HEK 293T, A549, HeLa (ATCC #CCL-2), MRC-5 (provided by D. Wang, Washington University), HFF-1 (ATCC #SCRC-1041), Hs 633T (Sigma-Aldrich #89050201), Huh7, RPE (provided by M. Mahjoub, Washington University), JEG3 (provided by I. Mysorekar, Washington University), U2OS (provided by S. Cherry, University of Pennsylvania), HT1080 (provided by J. Cooper, Washington University), Raji and K562 cells all were cultured at 37 °C in Dulbecco's Modified Eagle medium supplemented with 10% fetal bovine serum (FBS), 10 mM HEPES, 1 mM sodium pyruvate, 1× non-essential amino acids, and 100 U/ml of penicillin–streptomycin. HTR8/SV.neo cells (provided by I. Mysorekar, Washington University) were cultured in RPMI 1640 supplemented with 5% FBS and 1% penicillin and streptomycin. Jurkat cells were cultured at 37 °C in RPMI 1640 supplemented with 10% FBS and 10 mM HEPES. hCMEC/D3 cells (provided by R. Klein, Washington University) were cultured in EBM-2 medium (Lonza, 00190860) supplemented with 5% FBS, 5 μg/ml ascorbic acid, 10 mM HEPES, 1% lipid concentrate (Gibco, 11905-031) in plates pre-coated with collagen. All cell lines were tested and found to be free of mycoplasma contamination using a commercial kit. Cell lines were not authenticated.

Primary human keratinocytes (#102-05n), synovial fibroblasts (#408-05a), osteoblasts (#406-05f), chondrocytes (#402-05f) and skeletal muscle cells (#S150-05f) were purchased from Cell Applications. Primary human dermal fibroblasts (#CC-2509) were obtained from Lonza. Cells were thawed and cultured in specified medium according to the instructions of the manufacturers, and used within one week.

The following alphaviruses were used: CHIKV (strains 2006 La Reunion OPY1, 37997, AF15561, 181/25, and 181/25-mKate2 (rescued from pJM6 CHIKV-181/25 mKate2 cDNA clone, provided by T. Morrison and M. Heise), RRV (T48), MAYV (BeH407), ONNV (MP30), SFV (Kumba), SINV (Toto1101, Girdwood), Bebaru virus (BEBV, MM 2354), Middleburg virus (MIDV, 30037), Getah virus (GETV, AMM-2021), Una virus (UNAV, CoAr2380) and Barmah Forest virus (BRV, K10521). Additional viruses tested included chimaeric encephalitic alphaviruses (SINV–EEEV and SINV–WEEV), VEEV–GFP (TC-83)), a flavivirus (WNV, New York 2000), a bunyavirus (RVFV–GFP, MP-12), a rhabdovirus (VSV–GFP, Indiana) and a picornavirus (encephalomyocarditis virus, EMCV). Replication-competent SINV chimaeric viruses were constructed by replacing the SINV TR339 structural protein genes with EEEV FL93-939 or WEEV Fleming structural protein genes under control of the SINV subgenomic promoter in the TR339 cDNA clone[27]. All viruses were propagated in Vero cells and titrated by standard plaque or focus-forming assays[28].

Pooled sgRNA screen and data analysis. A GeCKOv2 CRISPR knockout pooled library encompassing 130,209 different sgRNAs against 20,611 mouse genes[29] was made available by F. Zhang (Addgene #1000000053), and amplified in Endura cells (Lucigen # 60242) as previously described[29,30]. The sgRNA library was divided in half (A + B), packaged into lentiviruses and used for independent screening. The sgRNA plasmid library was packaged in 293FT cells after co-transfection with psPAX2 (Addgene #12260) and pMD2.G (Addgene #12259) at a ratio of 2:2:1 using FugeneHD (Promega). Approximately 48 h after transfection, supernatants were collected, clarified by centrifugation (3,500 r.p.m. ×20 min) and aliquotted for storage at −80 °C.

For the CRISPR screen, a clonal 3T3-Cas9 cell line was generated by transduction with a packaged lentivirus (lentiCas9-Blast, Addgene #52962), blasticidin selection and limiting dilution. 3T3-Cas9 cells were expanded and transduced with CRISPR sgRNA lentivirus library at a multiplicity of infection (MOI) of 0.3 by spinoculation (1,000*g*) at 32 °C for 30 min in 12-well plates. After selection with puromycin for 7 days, $\sim 1 \times 10^8$ cells were inoculated with CHIKV-181/25-mKate2 (MOI of 1) and then incubated for 24 h to allow nearly all cells to become infected. Cells were sorted for an absence of mKate2 expression using a Sony Biotechnology Synergy SY3200 Cell sorter (Siteman Flow Cytometry Core, Washington University). To enrich for the cell population that was resistant to CHIKV infection and increase the signal-to-noise ratio, the mKate2-negative cells were expanded in culture. Given the rapid replication rate and cytopathic effect of CHIKV, two humanized neutralizing mAbs[5], CHK-152 and CHK-166 (2 μg/ml of each), were added to block infection by any residual virus. The expanded cells were re-infected with CHIKV-181/25-mKate2 in the absence of mAbs, sorted for mKate2 negative cells and the procedure was repeated for one additional round.

Genomic DNA was extracted from the uninfected cells (5×10^7) or the mKate2-negative sorted cells (1×10^7), and sgRNA sequences were amplified[31] and subjected to next generation sequencing using an Illumina HiSeq 2500 platform (Genome Technology Access Center, Washington University). The sgRNA sequences against specific genes were determined after removal of the tag sequences using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and cutadapt 1.8.1. sgRNA sequences were analysed using a published computational tool (MAGeCK)[6] (see Supplementary Tables 1, 2).

Gene validation. Mxra8 was validated by using three independent sgRNAs as follows: *Mxra8* sgRNA1, 5′-CTTGTGGATATGTATTCGGC-3′; *Mxra8* sgRNA2, 5′-TGTGCGCCTCGAGGTTACAG-3′; *Mxra8* sgRNA3, 5′-GCTGCATG ATCGCCAGCGCG-3′. The sgRNAs were cloned into the plasmid lentiCRISPR v.2 (Addgene #52961) and packaged with a lentivirus express system. 3T3 or MEF cells were transduced with lentiviruses expressing individual sgRNA and selected with puromycin for seven days before infection with different viruses. For some validation experiments, clonal cells edited by sgRNA1 were isolated by limiting dilution. To validate the human orthologue *MXRA8*, two different sgRNAs targeting all four isoforms were used: *MXRA8* sgRNA1, 5′-GGCGCGGATGCCTTTGAGCG-3′; *MXRA8* sgRNA2, 5′-GTCCGCCTGGAGGTCACCGA-3′. The sgRNAs were cloned similarly as above, and the gene-edited bulk cells were used for validation studies.

For flow cytometric analyses, gene-edited 3T3 cells were inoculated with different viruses as follows: CHIKV-181/25 (MOI of 3, 9.5 h), CHIKV-AF15561 (MOI of 10, 24 h), CHIKV-37997 (MOI of 3, 10 h), CHIKV-LR 2006 (MOI of 1, 9.5 h), ONNV (MOI of 3, 12 h), RRV (MOI of 3, 32 h), MAYV (MOI of 3, 24 h), SFV (MOI of 1, 9 h), SINV (MOI of 10, 6 h), SIN-WEEV (MOI of 10, 10 h), SIN-EEEV (MOI of 10, 10 h), VEEV–GFP (MOI of 3, 6.5 h), RVFV–GFP (MOI of 10, 8 h), VSV–GFP (MOI of 3, 6 h), WNV (MOI of 10, 25 h) and EMCV (MOI of 3, 6 h). Gene-edited MEF cells were inoculated with CHIKV-181/25 (MOI of 3, 8 h), CHIKV-AF15561 (MOI of 10, 10 h), CHIKV-37997 (MOI of 1, 10 h) and CHIKV-LR 2006 (MOI of 1, 8 h). Gene-edited MRC-5, RPE and HFF-1 cells were inoculated with CHIKV-181/25 (MOI of 10, 10 h), CHIKV-AF15561 (MOI of 10, 10 h) and CHIKV-LR 2006 (MOI of 1, 10 h). Gene-edited Hs 633 T cells were inoculated with CHIKV-181/25 (MOI of 15), CHIKV-AF15561 (MOI of 15) and CHIKV-LR 2006 (MOI of 3) for 11.5 h. At the indicated times, cells were collected with trypsin, and fixed with 1% paraformaldehyde (PFA) diluted in PBS for 15 min at room temperature and permeabilized with Perm buffer (Hank's Balanced Salt Solution (HBSS) supplemented with 10 mM HEPES, 0.1% (w/v) saponin, and 2% FBS) for 10 min at room temperature. Cells then were incubated for 30 min at room temperature with 1 μg/ml of the following virus-specific antibodies: CHIKV (mouse CHK-11[5]), ONNV and MAYV (mouse CHK-48), RRV (human 1I9); SINV (ascites fluid, ATCC VR-1248AF), SFV (mouse CHK-124), VEEV (mouse 1A4A), WEEV (mouse 11A1), EEEV (mouse EEEV-10), WNV (human E16[32]) and EMCV (mouse serum). After washing, cells were incubated with 2 μg/ml of Alexa Fluor 647-conjugated goat anti-mouse or anti-human IgG (Invitrogen) for 30 min at room temperature. Cells were processed on a MACSQuant Analyzer 10 (Miltenyi Biotec), and analysed using FlowJo software (Tree Star).

Validation also was performed by an infectious virus yield assay. Gene-edited 3T3 cells were plated 12 h before infection. Cells were inoculated with CHIKV-181/25, CHIKV AF15561, CHIKV-LR 2006 at an MOI of 0.01 or other alphaviruses (BEBV, MOI 0.001; BFV, GETV, UNAV, MIDV, and SFV, all at MOI of 0.01) for 1 h, then washed once and maintained in reduced 2% FBS culture medium. Supernatants were collected at specific times after infection for titration on Vero cells by focus-forming assay (CHIKV) or standard plaque assay (all other viruses).

Genomic RNA transfection and analysis. To assess for effects of Mxra8 on CHIKV replication, we transfected capped viral genomic RNA into MEF cells. Capped genomic RNA was generated using an mMESSAGE mMACHINE SP6 Transcription Kit according to the manufacturer's instructions (Thermo Fisher #AM1340) from the *NotI*-linearized CHIKV-181/25 cDNA clone. One microgram of purified RNA was transfected into control or Δ*Mxra8* cells using the Neon transfection system according to the manufacturer's instructions (Thermo Fisher Scientific). Cells were then incubated in 15 mM NH_4Cl to prevent subsequent rounds of infection. At specified times, cells were collected with trypsin and processed for E2 expression levels by flow cytometry.

Pseudotyped virus experiments. MLV–GFP pseudoviruses were made as described[33,34] except plasmids encoding structural proteins of CHIKV (strain 37997), VEEV (strain TrD) and EEEV (strain FL91-4697) were used. Pseudovirus entry in 3T3 cells expressing or lacking Mxra8 was assessed 36 h later by measuring GFP expression by flow cytometry.

Plasmid construction. The C-terminal Flag-tagged mouse Mxra8 corresponding to the transcript (NM_024263) was synthesized (Integrated DNA Technologies) and cloned into the lentivirus vector pCSII-EF1-IRES-Venus with restriction sties NotI/BamHI. The *Mxra8* sgRNA target sequences were mutated (cttgtggatatgtattcggcg to ctGgtCgaCatgtaCAGCgcg) to avoid re-cutting by Cas9 protein for the trans-complementation. Based on this plasmid, a truncation lacking the cytoplasmic domain was constructed by PCR-mediated mutagenesis, the ΔC-tail (378–442). To express the GPI-anchored Mxra8, the N-terminal 336 amino acids missing the transmembrane and cytoplasmic domains were fused with PLAP (ctggcgcccccgccggcaccaccgacgccgcgcacccggggcggtccgtggtccccgcgttgctcctctgctggccgggaccctgctgctggagacggccactgctccc) or the rodent herpesvirus Peru (RHVP) open reading frame R1 gene (tacccatacgatgttccagattacgctacgtcctcaccatccattggcggcccaaacatgactttactattggccatgatcatgtttgcgttaaagatagggtcg, HA tag

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

is underlined) GPI anchor sequences. To assess the function of different human MXRA8 isoforms, the cDNA of isoform 2 (NM_032348.3) containing C-terminal Myc and Flag tags was purchased from OriGene (Cat. No. RC200955), and cloned into the lentivirus vector pCSII-EF1-IRES-Venus. Isoform 1 (NM_001282585.1), isoform 3 (NM_001282584.1) and isoform 4 (NM_001282583.1) were created by either mutagenesis of isoform 2 or gene synthesis (Integrated DNA Technologies), and cloned into the lentivirus vector pCSII-EF1-IRES-Venus containing C terminus Myc and Flag tags.

Trans-complementation and ectopic expression experiments. The plasmids constructed above were packaged using the lentivirus expression system. Cells transduced with these lentiviruses were sorted for Venus-positive cells by flow cytometry. 3T3 cells were inoculated with CHIKV-181/25 (MOI 3) for 9.5 h or CHIKV-AF15561 (MOI 10) for 20 h. HeLa and A549 cells were inoculated with CHIKV-181/25 (MOI 3), CHIKV-AF15561 (MOI 10) or CHIKV-LR 2006 (MOI 1) for 14 h. 293T cells were inoculated with CHIKV-181/25 (MOI 1), CHIKV-AF15561 (MOI 3) or CHIKV-LR 2006 (MOI 0.5) for 11.5 h. CHO cells (K1 and 745) were inoculated with CHIKV-181/25 (MOI 0.3), CHIKV-AF15561 (MOI 10) or CHIKV-LR 2006 (MOI 0.3) for 12 h. Cells were then collected and processed for E2 expression by flow cytometry.

Generation and production of Mxra8–Fc and MXRA8-2–Fc. A cDNA fragment encoding the mouse Mxra8 extracellular domain (residues 22–336, GenBank accession number NM_024263) or the human MXRA8-2 extracellular domain (residues 20–337, GenBank accession number NM_032348.3) and the mouse IgG2b–Fc were synthesized (Integrated DNA Technologies) and inserted into the pCDNA3.4 vector (Thermo Fisher) downstream of an IL-2 signal peptide sequence. After confirmation by Sanger sequencing, Mxra8–Fc and MXRA8-2–Fc were expressed into Expi293 cells (Thermo Fisher). Cells were seeded at 5×10^6 cells per ml one day before transfection. Two hundred micrograms of plasmid was diluted in Opti-MEM (Thermo Fisher) and complexed with HYPE-5 transfection reagent before addition to cells. Transfected cells were supplemented daily with Expi293 medium and 2% (w/v) Hyclone Cell Boost. Four days after transfection, supernatant was collected, centrifuged at 3,000*g* for 15 min, and purified by protein A sepharose 4B (Thermo Fisher) chromatography. After elution and buffer neutralization, the purified protein was dialysed into 20 mM HEPES, 150 mM NaCl (pH 7.5), filtered through a 0.20-μm filter, and stored at −80 °C. Mxra8 that was cleaved from the IgG backbone was generated by inserting an HRV cleavage sequence (LEVLFQGP) into Mxra8–Fc downstream of the Mxra8 coding sequence and before the mouse IgG sequence. Mxra8-HRV–Fc was expressed in Expi293 cells as described above. After purification, Mxra8-HRV–Fc was cleaved using HRV 3C protease (Thermo Fisher) at a 1:10 ratio overnight at 4 °C. Cleaved Fc fragments were depleted using protein A sepharose chromatography, and the purity of Mxra8 was confirmed using SDS–PAGE analysis.

Expression of CHIKV virus-like particles. The CHIKV virus-like particles plasmid (strain 37997) was provided by G. Nabel (via the Vaccine Research Center, NIH) and expressed in Expi293 cells as previously described[18]. The supernatant was collected four days after transfection, 0.2-μm filtered and stored at 4 °C.

ELISA-based Mxra8–Fc binding assays. Anti-CHIKV human mAb 4N12[19] or anti-EEEV human mAb 53 (J.E.C., unpublished data) were immobilized (50 μl, 2 μg/ml) onto Maxisorp ELISA plates (Thermo Fisher) overnight in sodium bicarbonate buffer, pH 9.3. Plates were washed four times with PBS and blocked with PBS supplemented with 4% BSA for 1 h at room temperature. CHIKV-181/25 or SINV-EEEV was diluted to 1.5×10^7 FFU per ml in PBS and 50 μl per well was added for 1 h at room temperature. Mxra8–Fc and respective positive (CHK-11[5] and EEEV-10 (A.S.K. and M.S.D., unpublished results) and negative OPC–Fc (D.H.F., unpublished results) controls were diluted in PBS supplemented with 2% BSA and incubated for 1 h at room temperature. After serial washes with PBS, plates were incubated with horseradish peroxide conjugated goat anti-mouse IgG (H + L) (1:2000 dilution, Jackson ImmunoResearch) for 1 h at room temperature. After washing with PBS, plates were developed with 3,3′-5,5′ tetramethylbenzidine substrate (Dako) and 2N H_2SO_4. Absorbance was read at 450 nm with a TriStar Microplate Reader (Berthold). The mAb competition binding assay was performed by incubating 10 μg/ml of indicated anti-CHIKV human mAbs[19] for 30 min before the addition of mMxra8–Fc, as described above; the anti-CHIKV human mAbs were mapped previously to different epitopes by alanine scanning mutagenesis and evaluated for neutralizing activity[19]. Humanized anti-WNV mAb E16[32] was included as a negative control in competition binding assays.

Surface plasmon resonance based Mxra8 binding assay. Surface plasmon resonance binding experiments were performed on a Biacore T200 system (GE Healthcare) to measure the kinetics and affinity of Mxra8 binding to CHIKV virus-like particles. Experiments were performed at 30 μl/min and 25 °C using 0.01 M HEPES pH 7.4, 0.15 M NaCl, 3 mM EDTA, 0.005% v/v surfactant P20 as running buffer. CHK-265 mAb[5] was immobilized onto a CM5 sensor chip (GE Healthcare) using standard amine coupling chemistry, and CHIKV virus-like particles were captured. Recombinant Mxra8 was injected over a range of concentrations (1 μM to 20 nM) for 5 min followed by a 10 min dissociation period. As a negative control, murine norovirus was captured with mAb A6.2, as previously described[35]. Real-time data were analysed using BIAevaluation 3.1 (GE Healthcare). Kinetic profiles and steady-state equilibrium concentration curves were fitted using a global 1:1 binding algorithm with drifting baseline.

Cell-based Mxra8–Fc binding assay. 3T3 cells were inoculated with different viruses at the following MOI: CHIKV-181/25 (MOI of 3, 9.5 h), ONNV (MOI of 5, 12 h), MAYV (MOI of 3, 24 h), RRV (MOI of 3, 32 h), SINV (MOI of 10, 9 h) and VEEV (MOI of 3, 6.5 h). Cells were detached using TrypLE (Thermo Fisher #12605010), washed twice with cold HBSS supplemented with 15 mM HEPES and 2% FBS. Cells were incubated with 1 μg/ml of Mxra8–Fc, OPG–Fc or viral E2-specific antibodies at 4 °C for 25 min. After washing, cells were stained with goat anti-mouse IgG (H + L) conjugated with Alexa Fluor 647 for 25 min at 4 °C. After washing twice, cells were fixed with 2% PFA for 10 min at room temperature. After two additional washes, cells were subjected to flow cytometry analysis.

Virus binding and internalization assays. The assays were conducted in suspension. MEF cells were collected using TrypLE and washed twice with ice-cold medium supplemented with 2% FBS. CHIKV-AF15561 virions were purified through a 25% glycerol cushion at 25,000 r.p.m. for 2 h. For the binding assay, cells (5×10^5) and virions (MOI of 20) were mixed in a 1.5-ml microcentrifuge tube and incubated on ice for 45 min. After five cycles of centrifugation and washing, cells were lysed in RLT buffer for RNA extraction using an RNeasy Mini Kit (QIAGEN #74104). For the internalization assay, after 5 cycles of centrifugation and washing, cells were resuspended into medium supplemented with 2% FBS and 15 mM NH_4Cl and then incubated at 37 °C for 1 h. Cells were chilled on ice and treated with 500 ng/ml proteinase K on ice for 1 h. After three additional washes, cells were lysed in RLT buffer for RNA extraction. qRT-PCR was conducted with *Gapdh* as an internal control using a TaqMan RNA-to-CT 1-Step Kit (Thermo Fisher #4392938). Primers and probes used are as follows: Fwd CHK181/AF: 5′-TCGACGCGCCATCTTTAA-3′; rev CHK181/AF: 5′-ATCGAATGCACCGCACACT-3′; probe CHK181/AF: 5′ 6-FAM/ACCAGCCTG/ZEN/CACCCACTCCTCAGAC/3′ IABkFQ; fwd *Gapdh*: 5′-GTGGAGTCATACTGGAACATGTAG-3′; rev *Gapdh*: 5′-AATGGTGAAGGTCGGTGTG-3′; and probe *Gapdh*: 5′ 6-FAM/TGCAAATGG/ZEN/CAGCCCTGGTG/3′ IABkFQ.

For the flow cytometry-based binding assay, experiments were conducted as above but with 5×10^4 cells and an MOI of 200. After binding and washing, cells were fixed and stained with a mixture of mAbs (CHK-11, CHK-84, CHK-124 and CHK-166[5]) (1 μg/ml) at room temperature for 25 min. Cells were washed once and stained with 2 μg/ml of goat anti-mouse IgG (H + L) conjugated with Alexa Fluor 647 (Thermo Fisher #A21235) for 25 min. After two additional washes, cells were subjected to flow cytometry analysis.

Surface staining of mouse Mxra8 and human MXRA8. Mouse or human cells were collected with TrypLE and washed twice with cold HBSS supplemented with 15 mM HEPES and 2% FBS. Cells were incubated with anti-Mxra8 (mouse) Armenian hamster serum (1:300), anti-Mxra8 (mouse) hamster mAbs (1 μg/ml) or anti-MXRA8 (human) mAb (1 μg/ml) (MBL International # W040-3) at 4 °C for 25 min. After washing, cells were stained with 2 μg/ml goat anti-Armenian hamster IgG (H + L) conjugated with Alexa Fluor 647 (Abcam #ab173004) or goat anti-mouse IgG (H + L) conjugated with Alexa Fluor 647 (Thermo Fisher #A21235) for 25 min at 4 °C. After two additional washes, cells were fixed with 2% PFA for 10 min at room temperature. Cells were then washed twice and subjected to flow cytometry analysis.

Cell viability assay. A CellTiter-Glo Luminescent Cell Viability Assay (Promega) was performed according to the manufacturer's instructions. In brief, 2×10^4 3T3 or MEF cells in 100 μl culture medium were seeded into opaque-walled 96-well plates. 24 h later, 100 μl of CellTiter-Glo reagent was added to each well and allowed to shake for 2 min. After a 10-min incubation at room temperature, luminescence was recorded by using a Synergy H1 Hybrid Plate Reader (Biotek) with an integration time of 0.5 s per well.

Western blotting. Cells seeded in 6-well plates were washed once with PBS, chilled on ice in PBS and detached with a cell scraper. After spinning at 300 *g* for 5 min, cell pellets were lysed in 45 μl RIPA buffer (Cell Signaling #9806S) with a cocktail of protease inhibitors (Sigma-Aldrich # S8830). Samples were prepared in LDS buffer (Life Technologies) under reducing (+dithiothreitol) conditions. After heating (70 °C, 10 min), samples were electrophoresed using 10% Bis-Tris gels (Life Technologies) and proteins were transferred to PVDF membranes using an iBlot2 Dry Blotting System (Life Technologies). Membranes were blocked with 5% non-fat dry powdered milk and probed with hamster mAb 3G2.F5 (0.5 μg/ml) against mouse Mxra8. Western blots were developed using SuperSignal West Pico Chemiluminescent Substrate or SuperSignal West Femto Maximum Sensitivity Substrate (Life Technologies).

Alanine scanning mutagenesis for mapping. A CHIKV E2, 6K and E1 envelope protein expression construct (strain S27, Uniprot Reference #Q8JUX5) with

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

a C-terminal V5 tag was subjected to alanine scanning mutagenesis to generate a comprehensive mutation library[36]. Each residue of the envelope proteins was mutated to alanine, with alanine codons mutated to serine. One hundred and forty-one mutations within the E2 A and B domains were screened for binding by Mxra8–Fc. Binding of Mxra8–Fc to each mutant expressed in HEK-293T cells was determined by immunofluorescence detected with a high-throughput flow cytometer (Intellicyt), as previously described[36]. Residues of domains A or B were identified as contributing to the binding site if their mutation eliminated Mxra8–Fc binding, but supported binding of CHIKV mAbs that bind to the appropriate domain (control mAbs were CHKV-84, CHKV-88, IM-CKV063, IM-CKV065 and C9[5,36,37]).

Mapping of mutations onto the CHIKV p62–E1 crystal structure. Figures were prepared using the atomic coordinates of CHIKV p62–E1 (RCSB accession number 3N41) using the program PyMOL (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC).

mAb generation. Four-week-old male Armenian hamsters were immunized intravenously with 100 μg of purified Mxra8–Fc. After two boosts (~7 months of age), spleens were collected for hybridoma fusion and mAb production. Hybridoma supernatants were initially screened by ELISA using Mxra8-human Fc (mouse Fc was replaced by human Fc). As a second assay, we examined the binding of hybridoma supernatants to Mxra8 on the surface of 3T3 cells by flow cytometry. Finally, a tertiary screen evaluating blockade of CHIKV-181/25 infection by hybridoma supernatants was performed in 3T3 cells. After limiting dilution subcloning, the seven clones with the strongest blocking activities were selected and expanded. Antibodies were purified using protein A sepharose 4B (Invitrogen #101042), dialysed in PBS, concentrated and filtered for in vitro and in vivo experiments.

Blocking assays with Mxra8–Fc, MXRA8-2–Fc or anti-Mxra8 mAbs. Twenty-five thousand 3T3 or MRC-5 cells were seeded into 96-well plates 12 h before treatment. CHIKV-181/25 virions were purified through a 25% glycerol cushion at 25,000 r.p.m. for 2 h. Serially diluted Mxra8–Fc or MXRA8-2–Fc protein was incubated with purified virions (MOI of 3) for 1 h at 37 °C in a volume of 100 μl. The mixture was added to 3T3 or MRC-5 cells for 9.5 h or 11.5 h, respectively. Cells were then collected for intracellular E2 expression as measured by flow cytometry. For hamster mAb blocking experiments, 3T3 or MRC-5 cells were pre-incubated with serially diluted mAbs for 1 h at 37 °C in a volume of 50 μl, and then purified virions (MOI of 3) in 50 μl were added and incubated for 9.5 or 11.5 h, respectively. Cells were collected and intracellular E2 expression was analysed by flow cytometry. For hamster mAb blocking experiments on primary human cells (keratinocytes, dermal fibroblasts, synovial fibroblasts, osteoblasts, chondrocytes and skeletal muscle cells), 2×10^4 cells were seeded into 96-well plates 12 h before treatment. Cells were pre-incubated with Armenian hamster isotype control (Bio X Cell # BE0260), 1H1.F5 or 9G2.D6 mAb for 1 h at 37 °C in a volume of 50 μl (50 μg/ml) and, subsequently, purified CHIKV-AF15561 virions (MOI of 15) in 50 μl were added and incubated for 10.5 h. Cells were collected and intracellular E2 expression was analysed by flow cytometry.

Phosphatidylinositol-specific phospholipase C treatment. 3T3 cells (10^5) expressing GPI-anchored Mxra8 were collected using TrypLE and washed twice with PBS. Cells were then treated with 1 U/ml of phosphatidylinositol-specific phospholipase C (PI-PLC) (Sigma-Aldrich #P8804) in 50 μl PBS at 37 °C for 1 h. After two more cycles of washing, cells were stained for Mxra8 expression and processed by flow cytometry analysis as described above.

Mouse experiments. Experiments were carried out in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health after approval by the Institutional Animal Care and Use Committee at the Washington University School of Medicine (Assurance number A3381-01). Mxra8–Fc or an IgG control (JEV-13 mAb) (250 μg per mouse in PBS) was administered via an intraperitoneal route to four week-old wild-type male C57BL/6J mice 6 h before subcutaneous inoculation in the footpad with 10^3 FFU of CHIKV-AF15561. Alternatively, in co-injection experiments, CHIKV or ONNV was mixed directly with Mxra8–Fc or JEV-13 (25 μg per mouse in PBS), and incubated at 37 °C for 30 min before inoculation. At 12 h, 24 h and 72 h post-infection, animals were euthanized, and after perfusion with PBS, indicated tissues were collected. For antibody pre- or post-treatment experiments, 300 μg of purified hamster mAbs 1G11.E6 + 7F1.D8 or 4E7.D10 + 8F7.E1, and isotype control PIP (Bio X cell # BE0260) in PBS were administered via an intraperitoneal route to four week-old wild-type male C57BL/6 mice 12 h prior to, or 8 or 24 h after, subcutaneous inoculation in the footpad with 10^3 FFU of CHIKV-AF15561. Virus was titrated by focus forming assay as described[5] using mouse CHK-11[5] for CHIKV or mouse CHK-48 for ONNV as the detection antibody. Joint swelling was monitored at 72 h post-infection via left foot measurements (width × height) using digital calipers as previously described[38]. Samples sizes were estimated to determine a difference of three- to fivefold, depending on data distribution. Blinding and randomization were not performed.

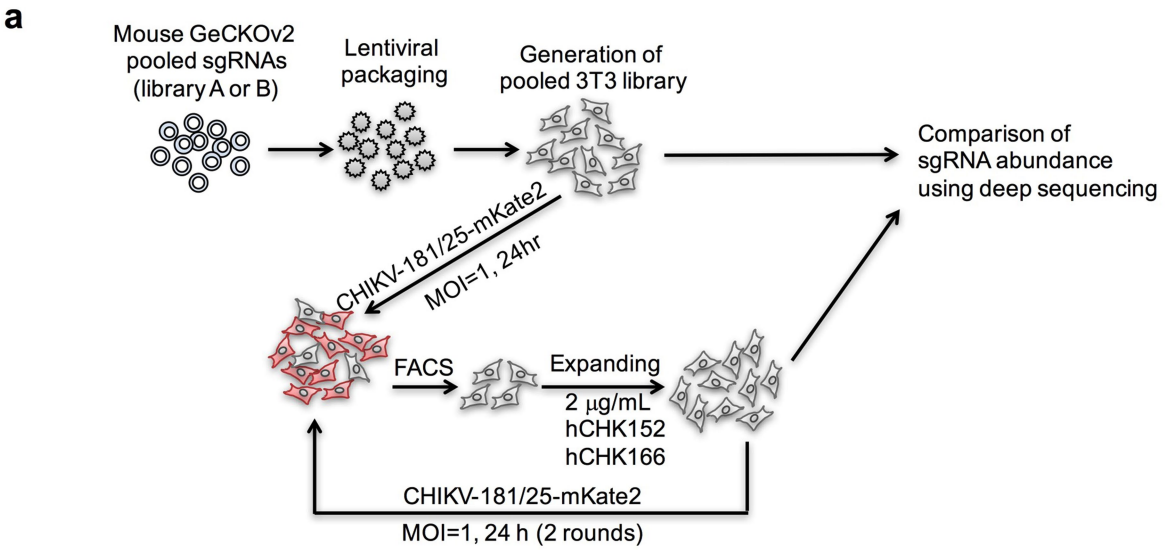
Statistical analysis. Statistical significance was assigned when P values were <0.05 using Prism Version 7 (GraphPad). Cell culture experiments were analysed by multiple t-tests with a Holm–Sidak correction or ANOVA with a multiple comparison correction. Analysis of levels of joint swelling or viral burden in vivo was determined by a Mann–Whitney, ANOVA, Kruskal–Wallis or unpaired t-test depending on data distribution and the number of comparison groups.

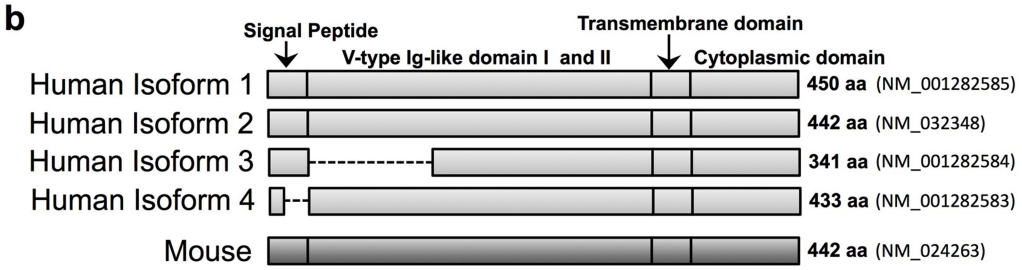
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

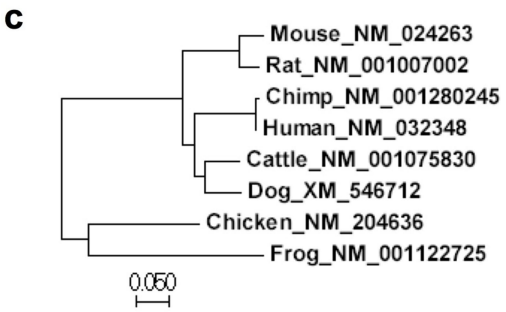
Data availability. All data supporting the findings of this study are available within the paper and its Supplementary Information. The Supplementary Tables provide data for the CRISPR–Cas9 screen, statistical analysis and alanine scanning mutagenesis mapping of the Mxra8–Fc binding site on CHIKV. All other data are available from the corresponding author upon reasonable request.

27. Klimstra, W. B., Ryman, K. D. & Johnston, R. E. Adaptation of Sindbis virus to BHK cells selects for use of heparan sulfate as an attachment receptor. *J. Virol.* **72**, 7357–7366 (1998).
28. Brien, J. D., Lazear, H. M. & Diamond, M. S. Propagation, quantification, detection, and storage of West Nile virus. *Curr. Protoc. Microbiol.* **31**, 15D.3.1–15D.3.18 (2013).
29. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* **11**, 783–784 (2014).
30. Joung, J. et al. Genome-scale CRISPR–Cas9 knockout and transcriptional activation screening. *Nat. Protocols* **12**, 828–863 (2017).
31. Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
32. Oliphant, T. et al. Development of a humanized monoclonal antibody with therapeutic potential against West Nile virus. *Nat. Med.* **11**, 522–530 (2005).
33. Huang, I. C. et al. SARS coronavirus, but not human coronavirus NL63, utilizes cathepsin L to infect ACE2-expressing cells. *J. Biol. Chem.* **281**, 3198–3203 (2006).
34. Jemielity, S. et al. TIM-family proteins promote infection of multiple enveloped viruses through virion-associated phosphatidylserine. *PLoS Pathog.* **9**, e1003232 (2013).
35. Taube, S. et al. High-resolution x-ray structure and functional analysis of the murine norovirus 1 capsid protein protruding domain. *J. Virol.* **84**, 5695–5705 (2010).
36. Selvarajah, S. et al. A neutralizing monoclonal antibody targeting the acid-sensitive region in chikungunya virus E2 protects from disease. *PLoS Negl. Trop. Dis.* **7**, e2423 (2013).
37. Jin, J. et al. Neutralizing monoclonal antibodies block chikungunya virus entry and release by targeting an epitope critical to viral pathogenesis. *Cell Reports* **13**, 2553–2564 (2015).
38. Hawman, D. W. et al. Chronic joint disease caused by persistent chikungunya virus infection is controlled by the adaptive immune response. *J. Virol.* **87**, 13878–13888 (2013).
39. Lubman, O. Y. et al. Rodent herpesvirus Peru encodes a secreted chemokine decoy receptor. *J. Virol.* **88**, 538–546 (2014).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.





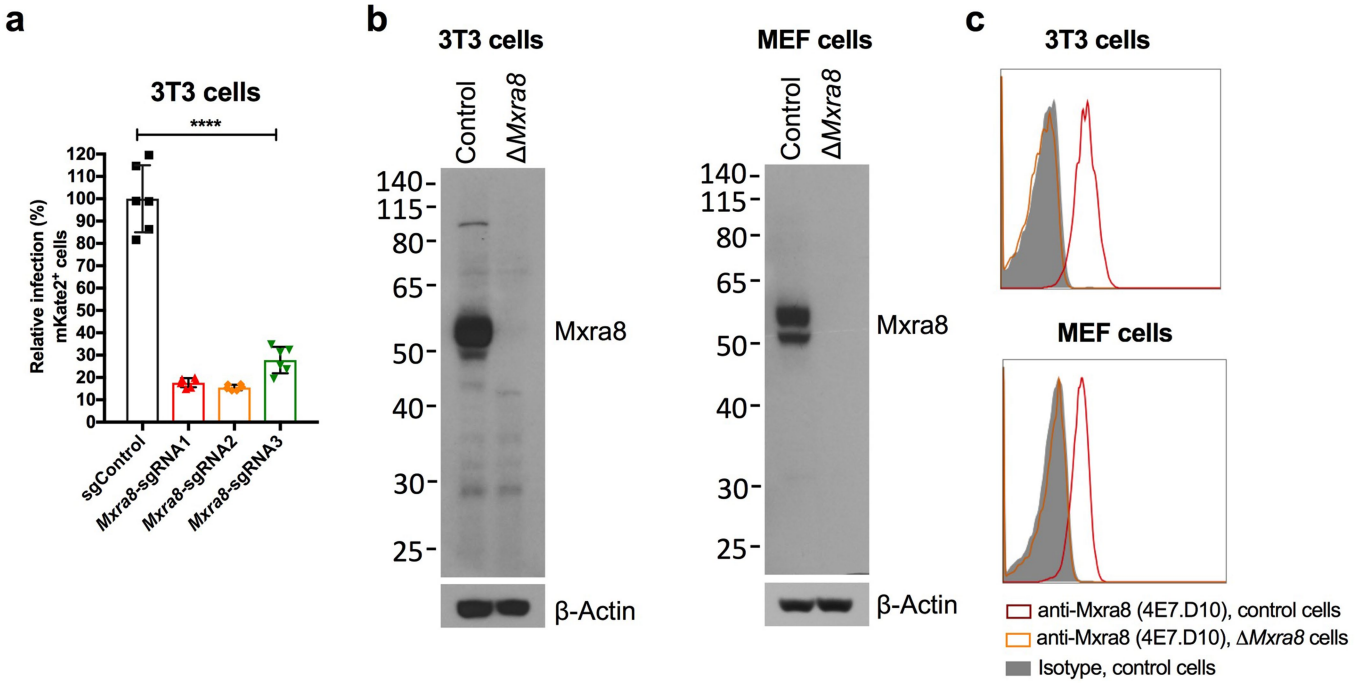


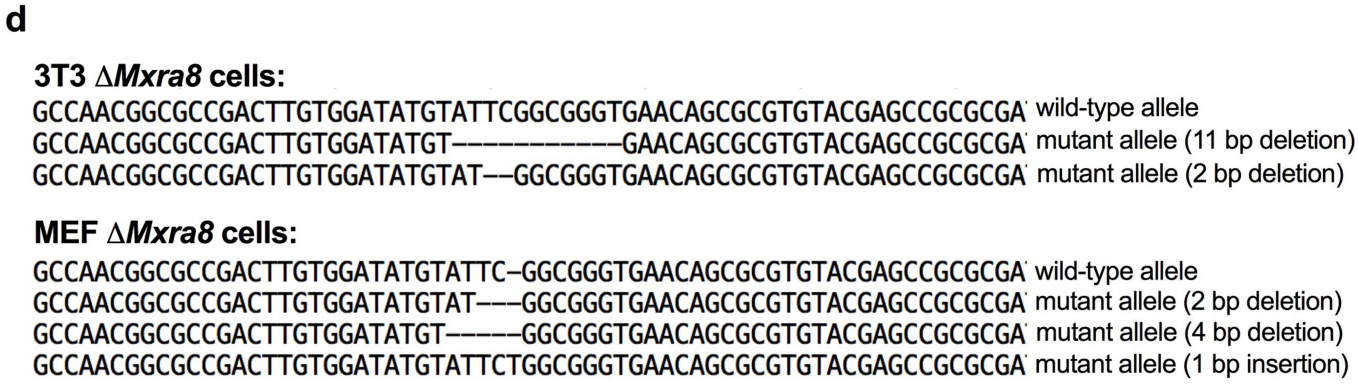
	Mouse	Rat	Chimp	Human	Cattle	Dog	Chicken	Frog
Mouse		80.5	77.4	78.1	76.5	79.8	58.9	52.9
Rat	83.5		67.7	68.2	66.2	69.9	50.8	45.8
Chimp	83.7	73.1		98.6	79.3	82.1	59.6	52.4
Human	84.4	73.5	98.6		80.0	82.8	60.0	52.8
Cattle	83.2	71.1	84.2	84.9		83.7	58.3	51.2
Dog	86.5	75.3	87.2	87.9	88.8		61.0	54.2
Chicken	75.3	65.2	74.2	75.3	72.6	75.3		63.2
Frog	69.9	61.6	67.2	68.1	65.9	68.5	78.0	

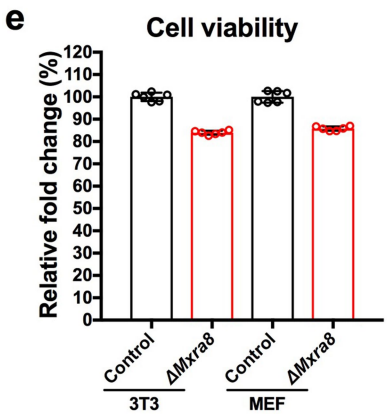
Identity
Similarity

Extended Data Fig. 1 | CRISPR–Cas9-based gene editing screen. **a**, Mouse 3T3 cells were transduced separately with two half libraries (A + B) comprising 130,209 sgRNAs, selected with puromycin and then inoculated with CHIKV-181/25-mKate2 (MOI of 1). One day later, mKate2-negative cells were sorted and then expanded in the presence of 2 μg ml^{-1} each of CHK-152 and CHK-166 neutralizing mAbs. Several days later, cells were re-inoculated with CHIKV-181/25-mKate2 without neutralizing mAbs and re-sorted for mKate2-negative cells. This procedure was repeated one additional time. Afterwards, genomic DNA was collected for sgRNA sequencing and compared to the parent library for abundance (see Supplementary Tables 1, 2). **b**, Diagram of the mouse Mxra8 and human MXRA8 orthologues. The transcript identification numbers and length of proteins are indicated to the right. Partial deletions in the isoforms 3 and 4 are shown as dashed lines. **c**, Phylogenetic tree of Mxra8 indicating genetic relationships. The neighbour-joining tree was constructed using MEGA 7 (https://www.megasoftware.net/). Scale bar shows the branch length. Right, identity (red) and similarity (yellow) matrix indicating the conservation of Mxra8 between species. The matrix was generated using MagGat 1.8 (http://iubio.bio.indiana.edu/soft/molbio/evolve/).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

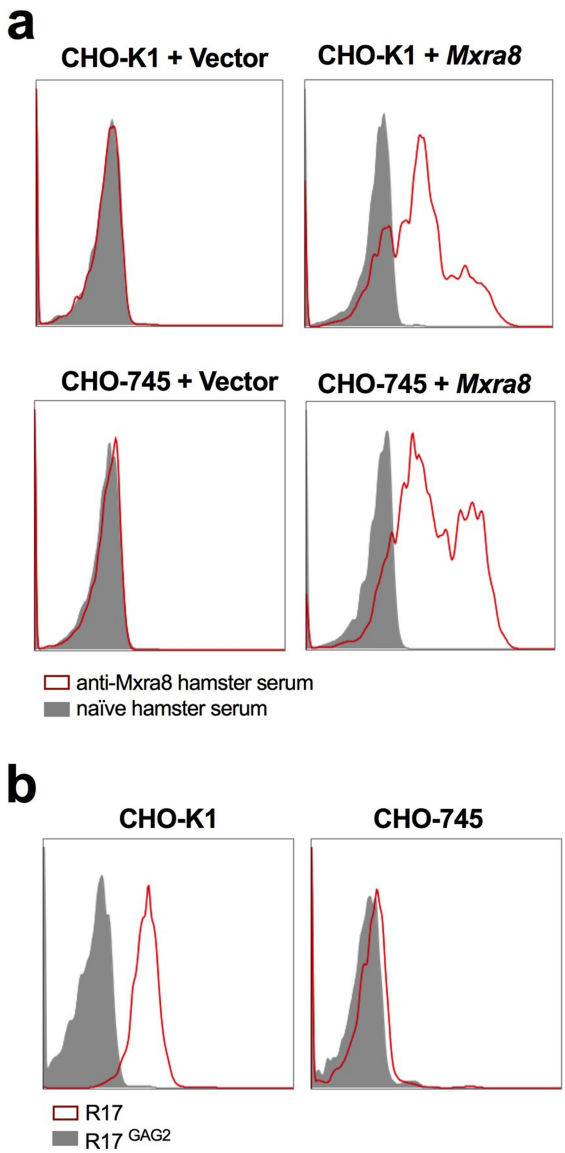


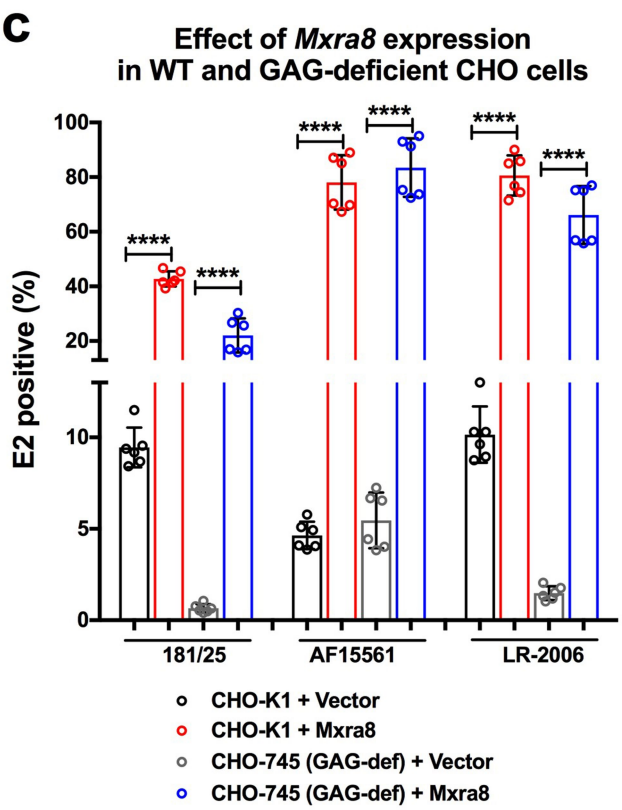




Extended Data Fig. 2 | Efficiency of targeting Mxra8 expression by CRISPR–Cas9 gene editing. a, 3T3 cells were edited with a control or three different *Mxra8* sgRNAs. After puromycin selection, bulk cells were inoculated with CHIKV-181/25-mKate2 and processed for marker gene expression by flow cytometry. Data are pooled from three experiments and expressed as mean ± s.d. ($n = 6$, one-way ANOVA with a Dunnett's multiple comparison test compared to control). **b**, Western blotting of Mxra8 in control and Δ*Mxra8* 3T3 or MEF cells using hamster mAb 3G2. F5. One representative of two is shown. **c**, 3T3 and MEF cells (parent or Δ*Mxra8*) were tested for Mxra8 surface expression by flow cytometry using anti-Mxra8 antibody (4E7.D10) and an isotype control mAb. One representative experiment of two is shown. **d**, Sanger sequencing of Mxra8 in control and Δ*Mxra8* 3T3 or MEF cells. Sequencing data shows an alignment and individual out-of-frame deletions. **e**, Viability of control and Δ*Mxra8* 3T3 and MEF cells. Equal numbers of cells were plated and viability was assessed over a 24-h period using the Cell-Titer Glo assay. The results were normalized to control cells and pooled from two experiments ($n = 6$). Error bars indicate s.d. ****$P < 0.0001$.

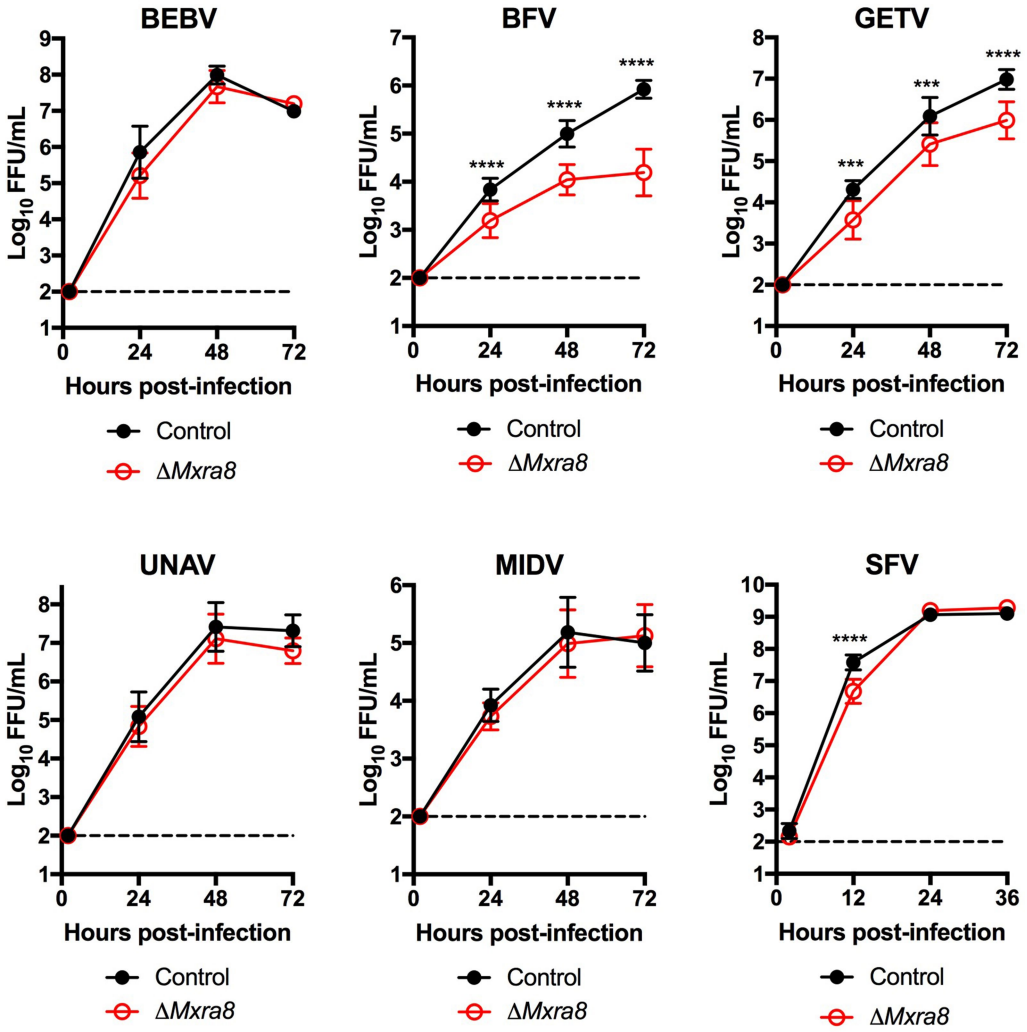
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.





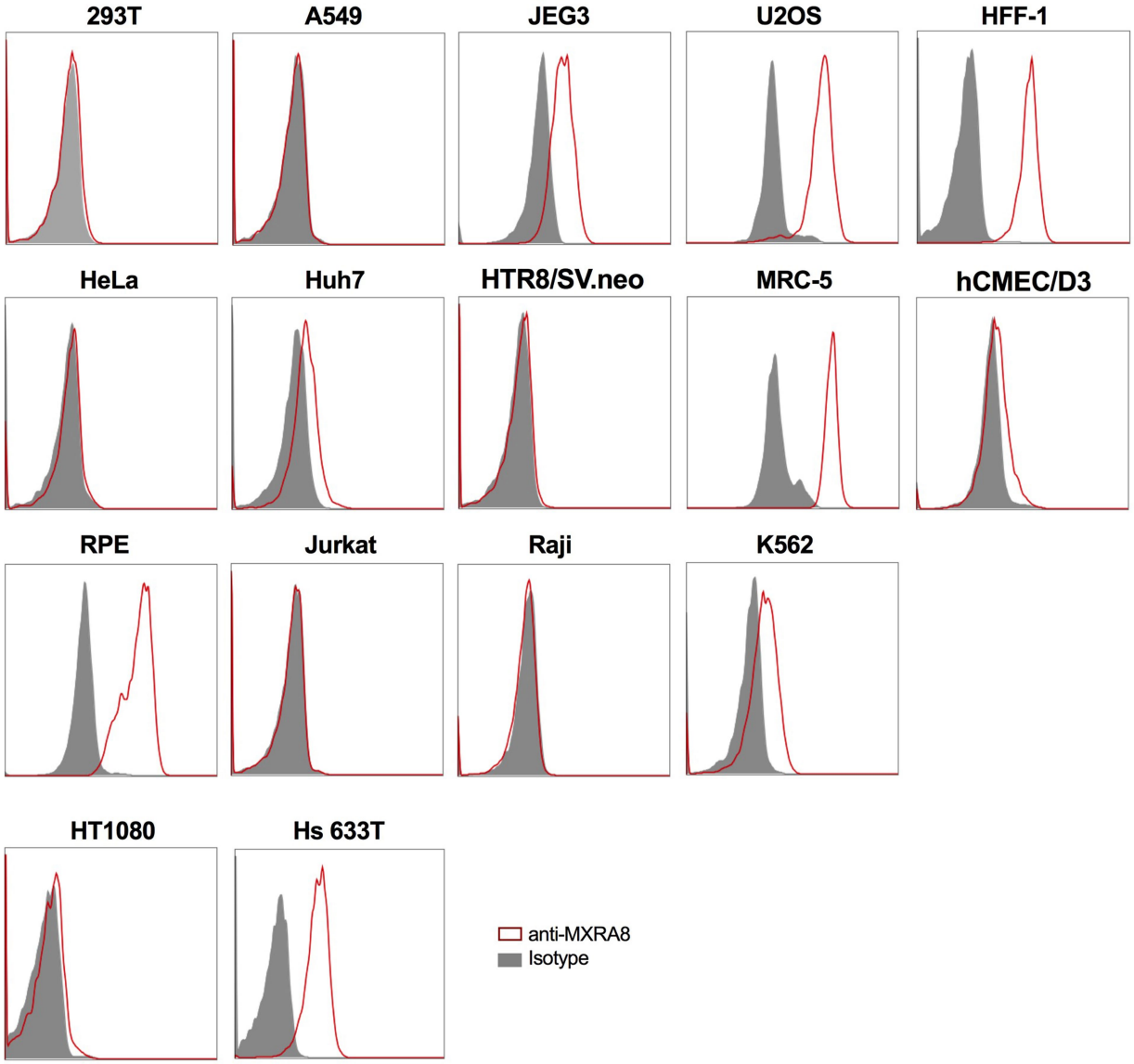
Extended Data Fig. 3 | CHIKV infectivity in CHO-K1 and CHO-745 cells in the presence or absence of ectopic Mxra8 expression. **a**, Surface expression of Mxra8 on CHO-K1 (wild type) and CHO-745 (glycosaminoglycan deficient[15]) cells stably transduced with control (vector-only) or mouse Mxra8 as judged by flow cytometry. **b**, Confirmation of heparan sulfate expression on the surface of CHO-K1 (wild type) and CHO-745 cells. Surface expression of heparan sulfate was evaluated using the R17 protein of rodent herpesvirus Peru, which binds to glycosaminoglycans on the surface of cells[39]. $R17^{GAG2}$ is a mutant form of the protein that lacks binding to glycosaminoglycans and served as a negative control. In **a** and **b**, data are representative of two experiments. **c**, CHO-K1 (wild type) and CHO-745 cells were transduced stably with control (vector) or mouse Mxra8 and inoculated with CHIKV (strains 181/25, AF15561 or LR-2006) and processed for intracellular E2 protein staining by flow cytometry. Data are from three experiments: mean ± s.d. ($n = 6$, one-way ANOVA with a Dunnett's multiple comparison test). $^{****}P < 0.001$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



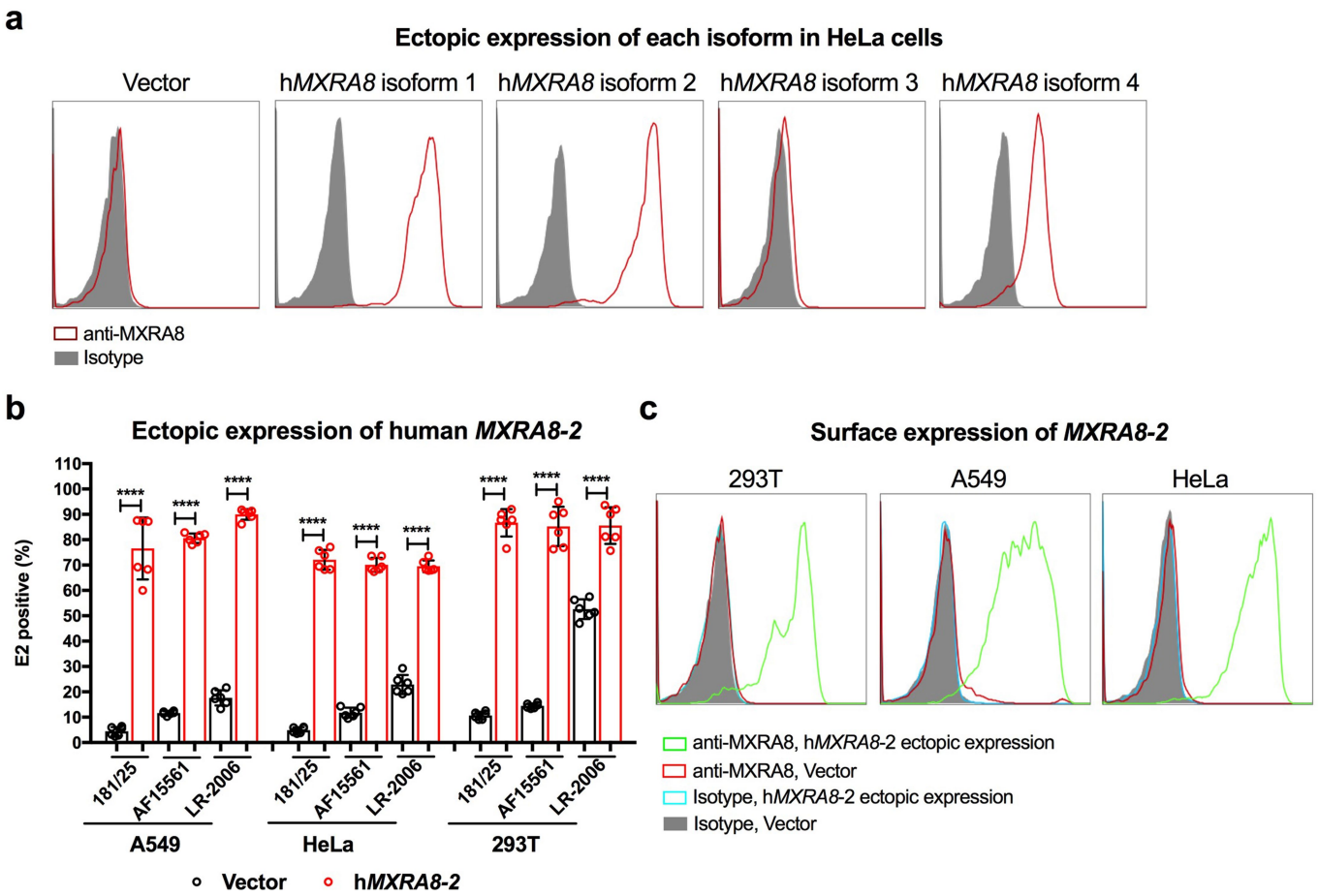
Extended Data Fig. 4 | Growth curves of related alphaviruses in Δ*Mxra8* 3T3 cells. Control and Δ*Mxra8* 3T3 cells were inoculated with BEBV, BFV, GETV, UNAV, MIDV or SFV at an MOI of 0.01 (except for BEBV, which was at 0.001), and supernatants were collected at the indicated times for focus forming assay. Data are pooled from two (BEBV) or three (all others) experiments and expressed as mean ± s.d. ($n = 6$, BEBV; $n = 9$, BFV, GETV, UNAV and MIDV; $n = 12$, SFV; two-way ANOVA with Sidak's multiple comparisons test). ***$P < 0.001$; ****$P < 0.0001$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



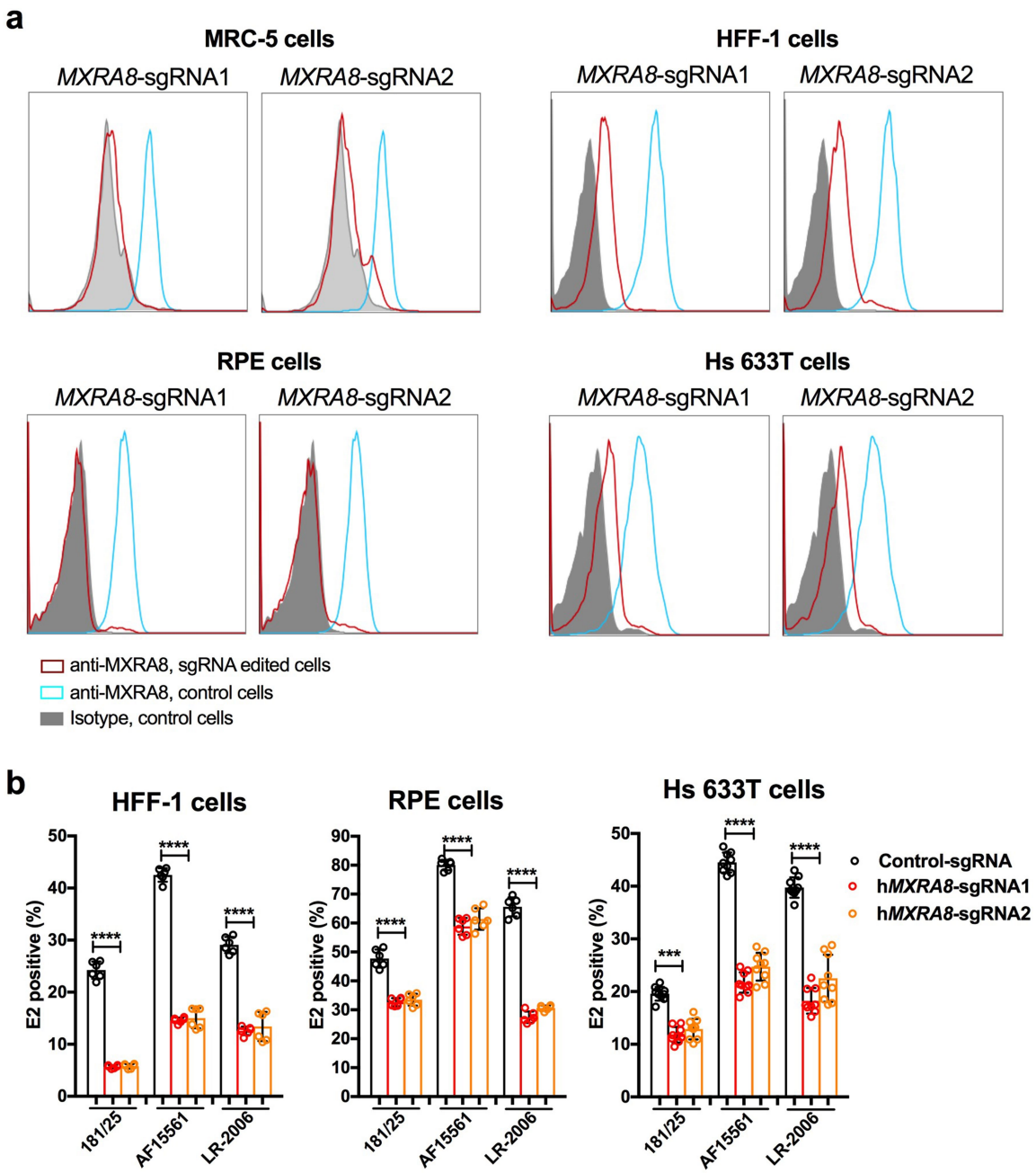
Extended Data Fig. 5 | Surface expression of MXRA8 in different human cell lines. Human cell lines were tested for MXRA8 surface expression by flow cytometry: 293T (embryonic kidney), A549 (lung adenocarcinoma), JEG3 (placental choriocarcinoma), U2OS (osteosarcoma), HFF-1 (foreskin fibroblasts), HeLa (cervical carcinoma), Huh7 (hepatocarcinoma), HTR8/SVneo (trophoblast progenitor), MRC-5 (lung carcinoma), hCMEC/D3 (cerebral microvascular endothelial cells), RPE (retinal pigment epithelial cell), Jurkat (T cell lymphoma), Raji (B cell lymphoma), K562 (eryrtholeukaemia), HT1080 (fibrosarcoma) and Hs 633T (fibrosarcoma) cells. Representative data are shown of two independent experiments. Grey histograms, isotype control mAb; red histograms, anti-MXRA8 mAb.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



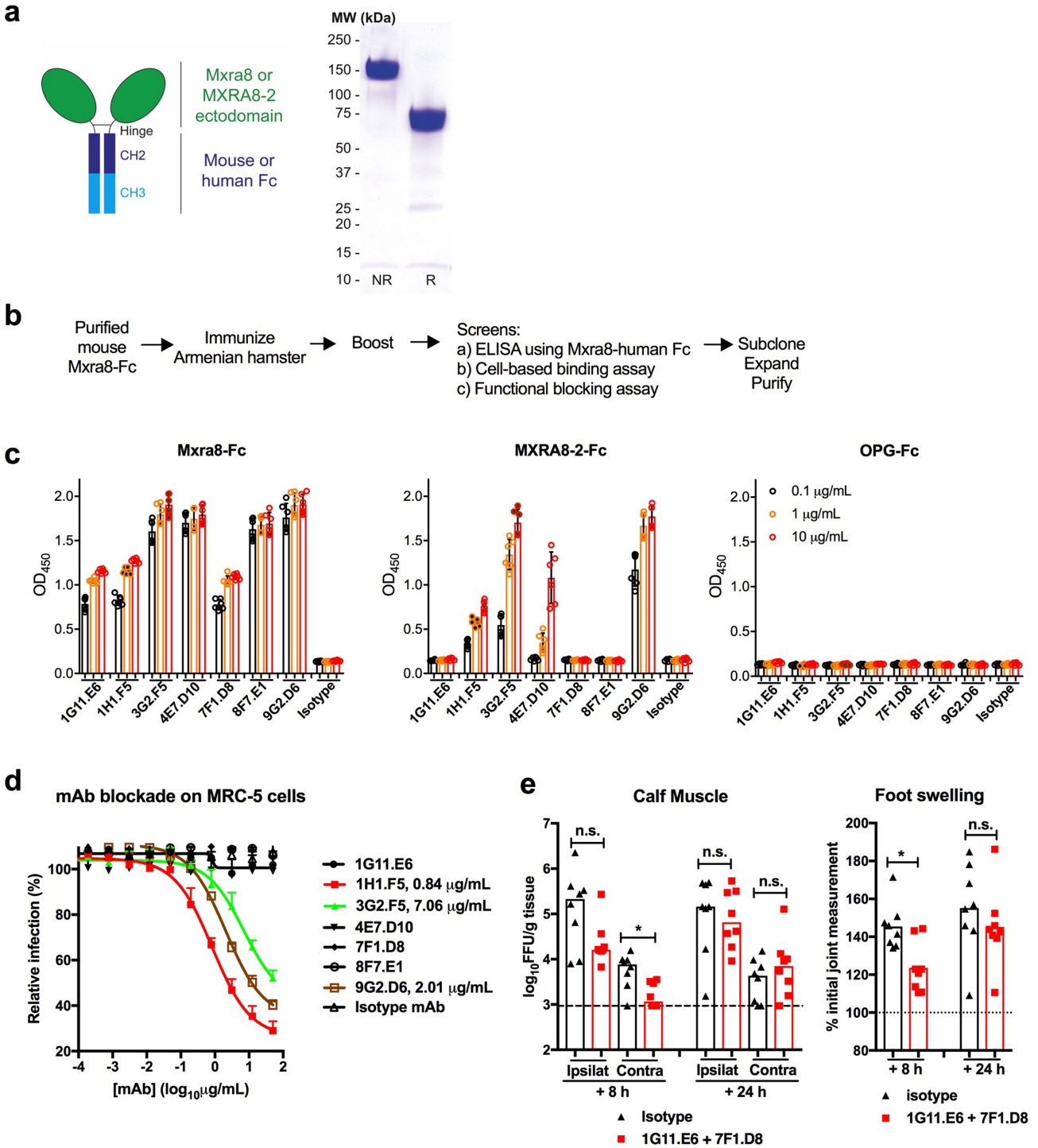
Extended Data Fig. 6 | MXRA8 supports enhanced infection of different CHIKV strains. a, Transduction and expression of different *MXRA8* (*MXRA8-1*, *MXRA8-2*, *MXRA8-3* and *MXRA8-4*) isoforms in HeLa cells. Representative data are shown from two experiments. Grey histograms, isotype control mAb; red histograms, anti-MXRA8 mAb. **b**, Effect of ectopic expression of *MXRA8-2* on CHIKV (strains 181/25, AF15561, and LR-2006) infection of A549, HeLa or 293T cells. Cells were collected and stained for CHIKV antigen with an anti-E2 antibody. Data are pooled from three experiments and expressed as mean ± s.d. ($n = 6$; two-tailed *t*-test with Holm–Sidak multiple comparison correction). **c**. Transduction and expression of *MXRA8-2* in 293T, A549, and HeLa cells. Representative data are shown from two experiments. ***$P < 0.001$; ****$P < 0.0001$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



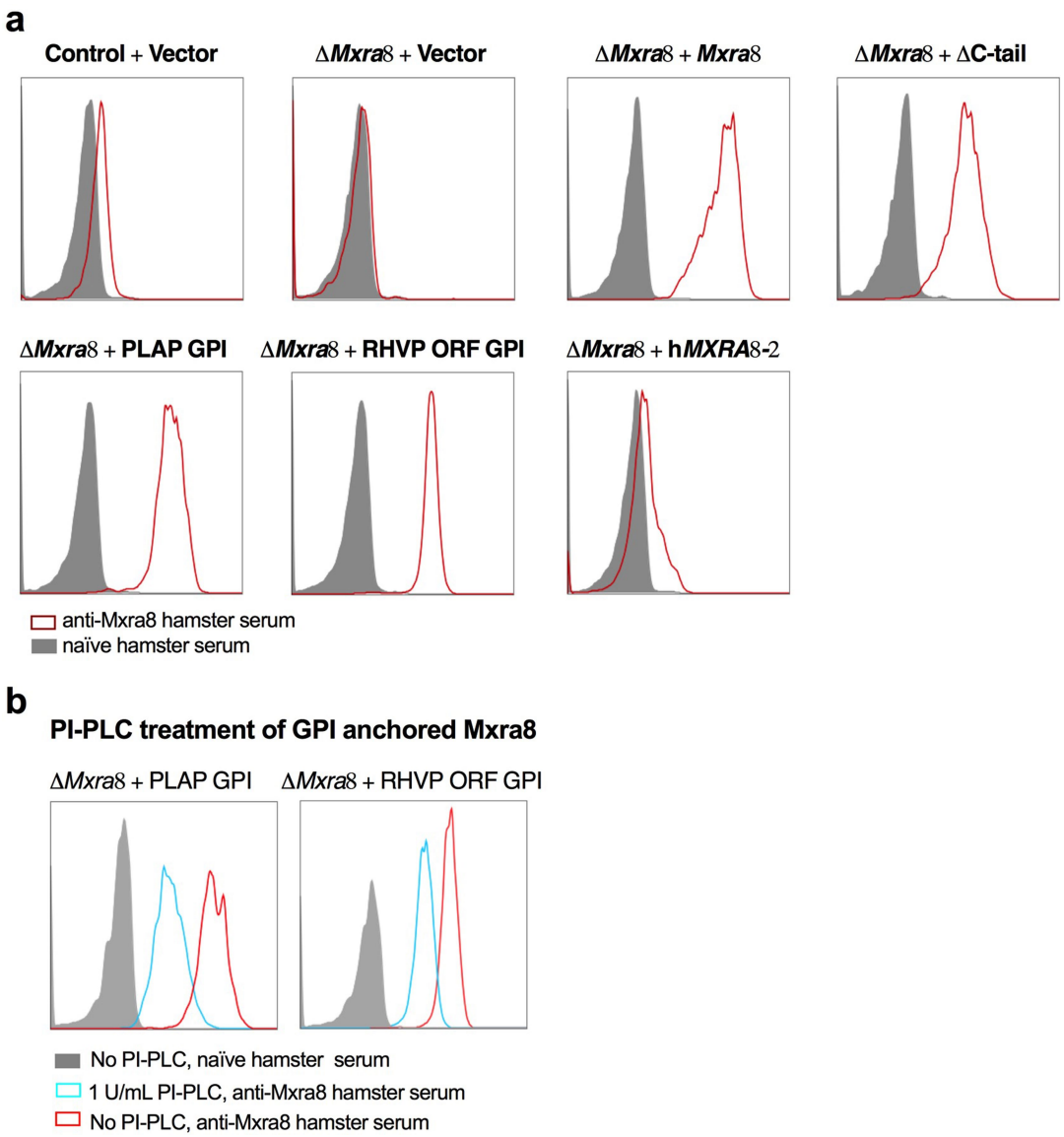
Extended Data Fig. 7 | Gene-editing of MXRA8 in human cell lines. **a**, Flow cytometry analysis of MXRA8 expression in human MRC-5, HFF-1, RPE and Hs 633T cells expressing control or two different *MXRA8* sgRNAs. Data are representative of two experiments. **b**, Gene-edited cells were inoculated with CHIKV (strains 181/25, AF15561 or LR-2006) in HFF-1, RPE, and Hs 633T cells. Cells were stained for viral antigen with an anti-E2 antibody. Data are pooled from two (HFF-1 and Hs 633T) or three (RPE) independent experiments and expressed as mean values ± s.d. ($n = 6$; one-way ANOVA with a Dunnett's multiple comparison test compared to the control). ****$P < 0.0001$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



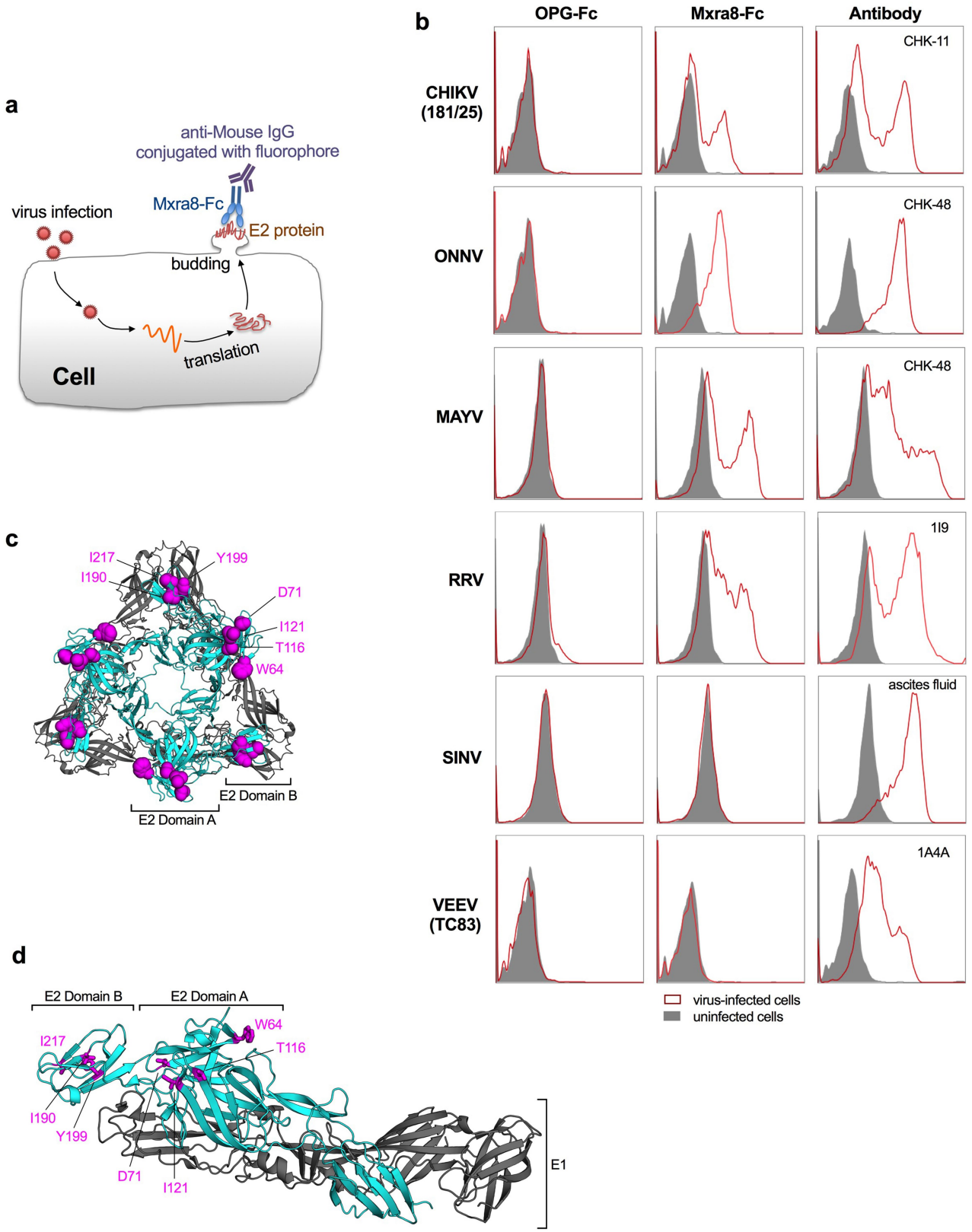
Extended Data Fig. 8 | Mxra8–Fc and anti-Mxra8 generation and function. a, Diagram of Mxra8–Fc (left) and SDS–PAGE (non-reducing (NR) and reducing (R) conditions) of purified material (right). Data are representative of three experiments. **b**, Scheme of anti-Mxra8 generation in Armenian hamsters. **c**, ELISA reactivity of anti-Mxra8 mAbs against Mxra8–Fc, MXRA8-2–Fc or OPG–Fc. Purified proteins (50 μl, 5 μg ml^{-1}) were immobilized overnight at 4 °C on ELISA plates. Anti-Mxra8 and isotype control mAbs were incubated for 1 h at room temperature. Signal was detected at 450 nm after incubation with horseradish peroxide conjugated goat anti-Armenian hamster IgG (H + L) and development with 3,3′-5,5′ tetramethylbenzidine substrate. **d**, Blockade of CHIKV-181/25 infection in MRC-5 cells with seven different hamster anti-Mxra8 or isotype control mAbs. mAbs were pre-incubated with cells for 1 h at 37 °C before addition of virus. After infection, cells were processed for E2 staining by flow cytometry. Relative infection was compared to a no mAb condition using flow cytometry and anti-E2 staining. Data in **c** and **d** are pooled from two experiments ($n = 6$) and expressed as mean ± s.d. **e**, Anti-Mxra8 mAbs (1G11 + 7F1) or isotype control hamster mAbs (300 μg total) were administered via intraperitoneal route 8 or 24 h after inoculation of CHIKV-AF15561 in the footpad. Left, at 72 h after initial infection, CHIKV titres were measured in the ipsilateral and contralateral gastrocnemius (calf) muscles. Right, at 72 h, ipsilateral foot swelling was measured and compared to measurements taken immediately before infection. Data are pooled from two experiments ($n = 8$; two-tailed Mann–Whitney test) and expressed as median values. *$P < 0.05$, n.s., not significant.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | Expression of truncated forms of Mxra8 mutants. a, Cell-surface expression of Mxra8 in Δ*Mxra8* 3T3 cells trans-complemented with vector, *Mxra8*, *Mxra8* ΔC-tail, *Mxra8* with GPI anchors (derived from PLAP or RHVP open reading frame R1) or *MXRA8-2*. Data are representative of two independent experiments. **b**, Effect of PI-PLC treatment on expression of different GPI-anchored forms or Mxra8. Data are representative of two independent experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Binding of Mxra8–Fc to surface-displayed E2 protein in virus-infected cells and mapping of Mxra8 binding site on E2. a, Diagram of the cell-based binding assay. After infection, viral structural proteins (for example, E2) traffic to the cell plasma membrane where progeny virion assembles and buds. E2 protein is displayed on the cell surface and is accessible to the binding of Mxra8–Fc and detection with a goat anti-mouse IgG secondary antibody by flow cytometry. **b**, Binding of Mxra8–Fc to virus-infected wild-type 3T3 cells. Cell were infected with the indicated viruses and processed for Mxra8–Fc binding by flow cytometry. Virus-specific anti-E2 antibodies were used as positive controls. Data are representative of two independent experiments. **c**, **d**, Mapped residues are shown as magenta spheres (**c**) or sticks (**d**) on the CHIKV p62–E1 structure (**c**, trimer of dimers, top view; **d**, heterodimer, side view) using PyMOL (Protein Data Bank code: 3N41). The E1 and E2 proteins are coloured in grey and cyan, respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0130-2

Bystander CD8$^+$ T cells are abundant and phenotypically distinct in human tumour infiltrates

Yannick Simoni[1]*, Etienne Becht[1], Michael Fehlings[1,2], Chiew Yee Loh[1], Si-Lin Koo[3], Karen Wei Weng Teng[1], Joe Poh Sheng Yeong[1,4], Rahul Nahar[5], Tong Zhang[5], Hassen Kared[1], Kaibo Duan[1], Nicholas Ang[1], Michael Poidinger[1], Yin Yeng Lee[5], Anis Larbi[1], Alexis J. Khng[5], Emile Tan[6], Cherylin Fu[6], Ronnie Mathew[6], Melissa Teo[7], Wan Teck Lim[3], Chee Keong Toh[3], Boon-Hean Ong[8], Tina Koh[7], Axel M. Hillmer[5], Angela Takano[4], Tony Kiat Hon Lim[4,5,9], Eng Huat Tan[3], Weiwei Zhai[5], Daniel S. W. Tan[3,5], Iain Beehuat Tan[3,5,9] & Evan W. Newell[1]*

Various forms of immunotherapy, such as checkpoint blockade immunotherapy, are proving to be effective at restoring T cell-mediated immune responses that can lead to marked and sustained clinical responses, but only in some patients and cancer types[1–4]. Patients and tumours may respond unpredictably to immunotherapy partly owing to heterogeneity of the immune composition and phenotypic profiles of tumour-infiltrating lymphocytes (TILs) within individual tumours and between patients[5,6]. Although there is evidence that tumour-mutation-derived neoantigen-specific T cells play a role in tumour control[2,4,7–10], in most cases the antigen specificities of phenotypically diverse tumour-infiltrating T cells are largely unknown. Here we show that human lung and colorectal cancer CD8$^+$ TILs can not only be specific for tumour antigens (for example, neoantigens), but also recognize a wide range of epitopes unrelated to cancer (such as those from Epstein–Barr virus, human cytomegalovirus or influenza virus). We found that these bystander CD8$^+$ TILs have diverse phenotypes that overlap with tumour-specific cells, but lack CD39 expression. In colorectal and lung tumours, the absence of CD39 in CD8$^+$ TILs defines populations that lack hallmarks of chronic antigen stimulation at the tumour site, supporting their classification as bystanders. Expression of CD39 varied markedly between patients, with some patients having predominantly CD39$^-$ CD8$^+$ TILs. Furthermore, frequencies of CD39 expression among CD8$^+$ TILs correlated with several important clinical parameters, such as the mutation status of lung tumour epidermal growth factor receptors. Our results demonstrate that not all tumour-infiltrating T cells are specific for tumour antigens, and suggest that measuring CD39 expression could be a straightforward way to quantify or isolate bystander T cells.

Using mass cytometry and a panel dedicated to the detailed profiling of tumour infiltrating T cells we observed that, consistent with previous reports[5,6,11], CD8$^+$ TILs constitute a highly heterogeneous cell population both within individual tumours (Fig. 1a) and among patients with lung and colorectal tumours ($n = 144$ patient tumours analysed by mass cytometry in this study) (Fig. 1b and Extended Data Fig. 1). We therefore decided to investigate the antigen specificity of CD8$^+$ TILs to better understand the basis for this heterogeneity. In total, we screened for 1091 putative neoantigens, 123 tumour-associated antigens (TAA) and 46 cancer-unrelated epitopes (mostly virus-derived) using mass cytometry coupled to multiplex major histocompatibility complex (MHC)-tetramer staining, as reported previously[12] (Fig. 2, Extended Data Fig. 2 and Supplementary Tables 1–3). Two positive hits were detected for neoantigen epitopes from a total of 24 patients tested (Fig. 2b, c and Supplementary Table 4). As 0.18% of the 1,091 computationally-predicted putative neoantigens could be confirmed experimentally, these data are in line with other publications reporting identification rates for neoantigen-specific CD8$^+$ T cells from predicted neoantigens of between 0% and 0.5%[13–15]. The small number of neoantigen-specific T cell populations detected may also be related to the relatively low mutational burden of these tumours (Supplementary Table 4), even though neoantigen-specific T cell responses have previously been reported in the context of other tumours with low mutational burden[13,16]. Nevertheless, these results highlight the challenge of accurately predicting and validating neoantigens for therapeutic purposes[9]. We also detected two tumour-specific CD8$^+$ TIL populations in an unusual case of lung cancer associated with Epstein–Barr virus (EBV) infection (lymphoepithelioma-like carcinoma, LELC) (Extended Data Fig. 3). Despite testing 40 patient tumours with large panels of TAA-derived epitopes, we failed to identify TAA-specific CD8$^+$ TILs. Data from in vitro expanded CD8$^+$ T cells have shown that these cells can be detected by MHC-tetramer staining[17–19]. MART-1 and NY-ESO-1 epitope-specific T cells have also been detected in unexpanded TILs[20,21]. It is possible that TAA-specific CD8$^+$ TIL cells were absent or present at undetectably low frequencies in all of the samples we tested.

Unexpectedly, we detected cancer-unrelated MHC-tetramer$^+$ cells ($n = 46$ CD8$^+$ T cell populations) in cohorts of patients with lung cancer or colorectal cancer (in 9 of 24 lung cancer patients, 37.5%; and in 21 of 42 colorectal cancer patients, 50%) (Fig. 2b). In these cases, MHC-tetramer$^+$ CD8$^+$ TILs were specific for various Epstein Barr virus (EBV), human cytomegalovirus (HCMV) or influenza virus epitopes that were presented by three different HLA alleles (Fig. 2d). Frequencies for individual epitopes varied between 0.07% and 3.3% of total CD8$^+$ TILs, and 11 examples of these were validated using fluorescence flow cytometry (Fig. 2d, e, Extended Data Fig. 4). The expression of CD69 and/or CD103 in many of these cancer-unrelated CD8$^+$ TILs suggests that they are not derived from blood contamination (Extended Data Fig. 4). These data therefore show that CD8$^+$ TILs are not all specific for tumour antigens, but can include bystander CD8$^+$ TILs that are specific for cancer-unrelated epitopes.

Having identified cancer-unrelated bystander and tumour-specific CD8$^+$ TILs, we next compared the phenotypes of these two populations with those of remaining CD8$^+$ TILs of unknown specificity. All the tumour-specific CD8$^+$ TILs that we identified displayed resident memory T cell-like phenotypes and expressed various co-stimulatory and inhibitory receptors, such as PD-1 (Fig. 3a and Extended Data Fig. 3, 5). Surprisingly, we observed overlapping but diverse phenotypic profiles for cancer-unrelated CD8$^+$ TILs with respect to these markers. Many of the bystander CD8$^+$ TILs also expressed resident memory T cell-like phenotypes as well as various co-stimulatory

[1]Agency for Science, Technology and Research (A*STAR), Singapore Immunology Network (SIgN), Singapore, Singapore. [2]immunoSCAPE, Singapore, Singapore. [3]Division of Medical Oncology, National Cancer Centre Singapore (NCCS), Singapore, Singapore. [4]Department of Anatomical Pathology, Singapore General Hospital, Singapore, Singapore. [5]Agency for Science, Technology and Research (A*STAR), Genome Institute of Singapore (GIS), Singapore, Singapore. [6]Department of Colorectal Surgery, Singapore General Hospital, Singapore, Singapore. [7]Division of Surgical Oncology, National Cancer Centre Singapore (NCCS), Singapore, Singapore. [8]Department of Cardiothoracic Surgery, National Heart Centre Singapore (NHCS), Singapore, Singapore. [9]Duke–National University of Singapore Medical School, Singapore, Singapore. *e-mail: yannick_simoni@immunol.a-star.edu.sg; evan_newell@immunol.a-star.edu.sg

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

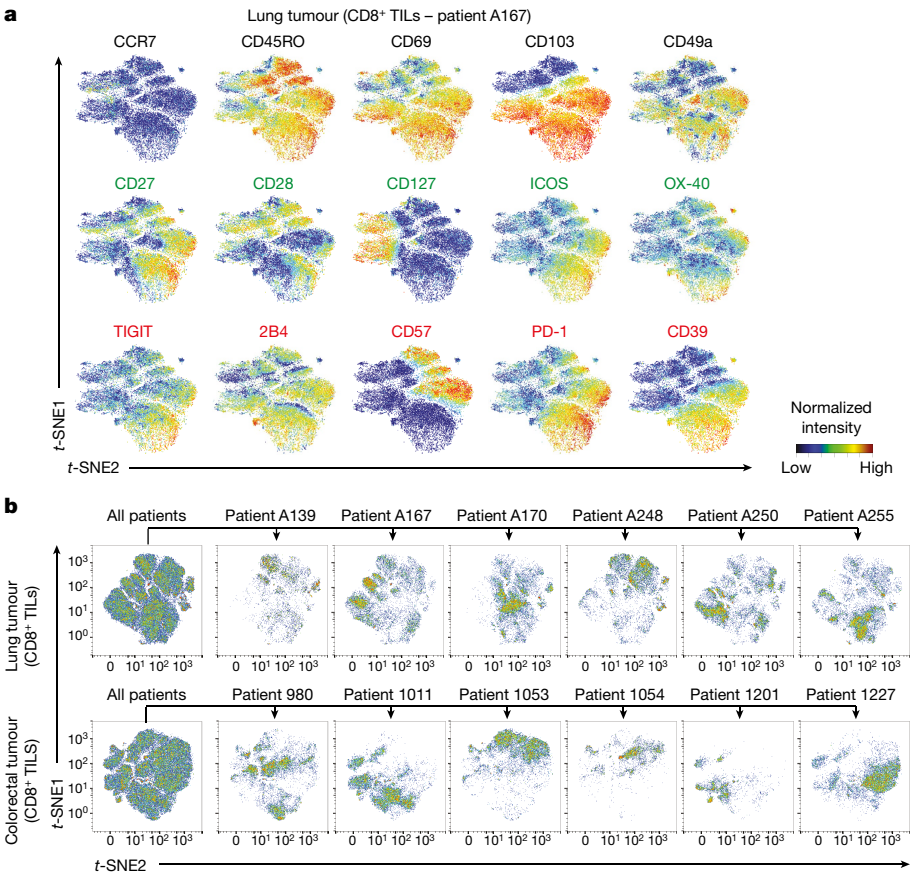


Fig. 1 | Tumour-infiltrating CD8$^+$ TILs are phenotypically heterogeneous within a tumour and across patients. a, *t*-distributed stochastic neighbour embedding (*t*-SNE) map of CD8$^+$ TILs isolated from a colorectal tumour. *t*-SNE was performed on one patient to explore the heterogeneity of CD8$^+$ TILs within an individual (see also Extended Data Fig. 1 and Methods). Representative data from one patient (see Methods for source data availability). **b**, *t*-SNE map of CD8$^+$ TILs isolated from lung tumours or colorectal tumours. Mass cytometry and *t*-SNE were performed simultaneously on six different patients from each cohort to explore the heterogeneity of CD8$^+$ TILs across patients. Patient identifiers refer to individual patients. Representative data from $n = 6$ patients for each cancer type.

and activation marker molecules. The inhibitory receptors TIGIT and PD-1, two markers that were previously shown to be expressed by tumour-antigen-specific CD8$^+$ T cells[22], were also expressed by many of these cells (Fig. 3a, c). Although PD-1 has been proposed as a marker of tumour-specific CD8$^+$ T cells[23], our results are consistent with previous reports of virus-specific CD8$^+$ T cells infiltrating tumours that express PD-1 in mice[24]. However, we observed a striking lack of CD39 expression in bystander CD8$^+$ TILs (5.2 ± 8.4% (s.d. is used throughout), $n = 46$). By contrast, CD39 was highly expressed by tumour-specific CD8$^+$ TILs and variably expressed by cells of unknown specificity (40.4 ± 27.2%, $P < 0.0001$) (Fig. 3b, c). CD39 is a transmembrane extracellular ATPase that is widely expressed by regulatory T cells, B cells and some tumour cells. In conjunction with the enzymatic activity of CD73, CD39 can catalyse the conversion of ATP to adenosine, which has been shown to have immunosuppressive activity[25,26]. Based on these data, we hypothesize that the lack of CD39 could be used to enrich for cancer-unrelated bystander CD8$^+$ TILs. Conversely, though we think that these results suggest that CD39 could also be a useful marker of tumour-specific CD8$^+$ TILs, this link could be observed only in two neoantigen responses from two patients and two tumour-specific CD8$^+$ TIL populations in an unusual LELC tumour.

To better compare the characteristics of CD39$^-$ CD8$^+$ and CD39$^+$ CD8$^+$ TILs, we performed transcriptomic profiling. Using principal component analysis (PCA) and gene set enrichment analysis (GSEA), we found that CD39$^+$ CD8$^+$ TILs were enriched in expression of genes related to cell proliferation and exhaustion, which are characteristics of chronically stimulated T cells[27–29] (Fig. 4a, b and Extended Data Fig. 6), consistent with previous reports in both cancer[25] and infectious disease[26]. In line with this, T cell receptor (TCR) sequencing indicated a skewed and reduced diversity of TCR sequence diversity in CD39$^+$ CD8$^+$ TILs (Extended Data Fig. 7), supporting the notion of enrichment for cells that have undergone tumour-antigen-driven clonal expansion[29,30]. At the protein level, compared to their CD39$^-$ counterparts, CD39$^+$ CD8$^+$ TILs from colon and lung tumours showed hallmarks of exhausted cells in terms of both phenotypic and functional markers (Fig. 4c, d and Extended Data Fig. 8), consistent with the transcriptomic profiling data. Thus, expression of CD39 defines a population of highly exhausted cells; whereas the absence of CD39 in CD8$^+$ TILs defines a population whose phenotype is inconsistent with chronic antigen stimulation at the tumour site, consistent with a bystander role.

Next, we investigated whether expression of CD39 by CD8$^+$ TILs was linked to clinical parameters measured in either of the studied patient cohorts. In colorectal tumours, we detected highly heterogeneous frequencies of CD39 expression among CD8$^+$ TILs ($n = 94$; mean, 44.5 ± 23.7%; minimum, 0.2%; maximum, 85.8%) (Fig. 4e and Extended Data Fig. 9). No significant correlations with any clinical parameters (tumour mutational burden, driver mutational status or consensus molecular subtype (CMS)) were obtained from this cohort (Extended Data Fig. 9). However, based on the transcriptomic profiles of adjacent frozen tumour sections, we found that tumours with higher percentages of CD39$^+$ CD8$^+$ TILs had gene expression profiles indicative of T cell inflammation and expression of pathways associated with antigen processing and presentation (Extended Data Fig. 10). In lung cancer patients, the expression of CD39 in CD8$^+$ TIL cells was also highly heterogeneous across patients ($n = 50$; mean, 21.9 ± 23.23%; minimum, 0%; maximum, 88.5%) (Fig. 4e and Extended Data Fig. 9). For this type of cancer, the frequency of CD39$^+$ CD8$^+$ TIL cells clearly correlated with the epidermal growth factor receptor (EGFR)-mutation status, an oncogenic driver mutation that is especially common in East Asian patients with lung cancer[31]. Preliminary reports suggest that patients with EGFR-mutated tumours are relatively poor responders to anti-PD-1 treatment and have low CD8$^+$ T cell density compared to patients with EGFR-wild-type tumours[32].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

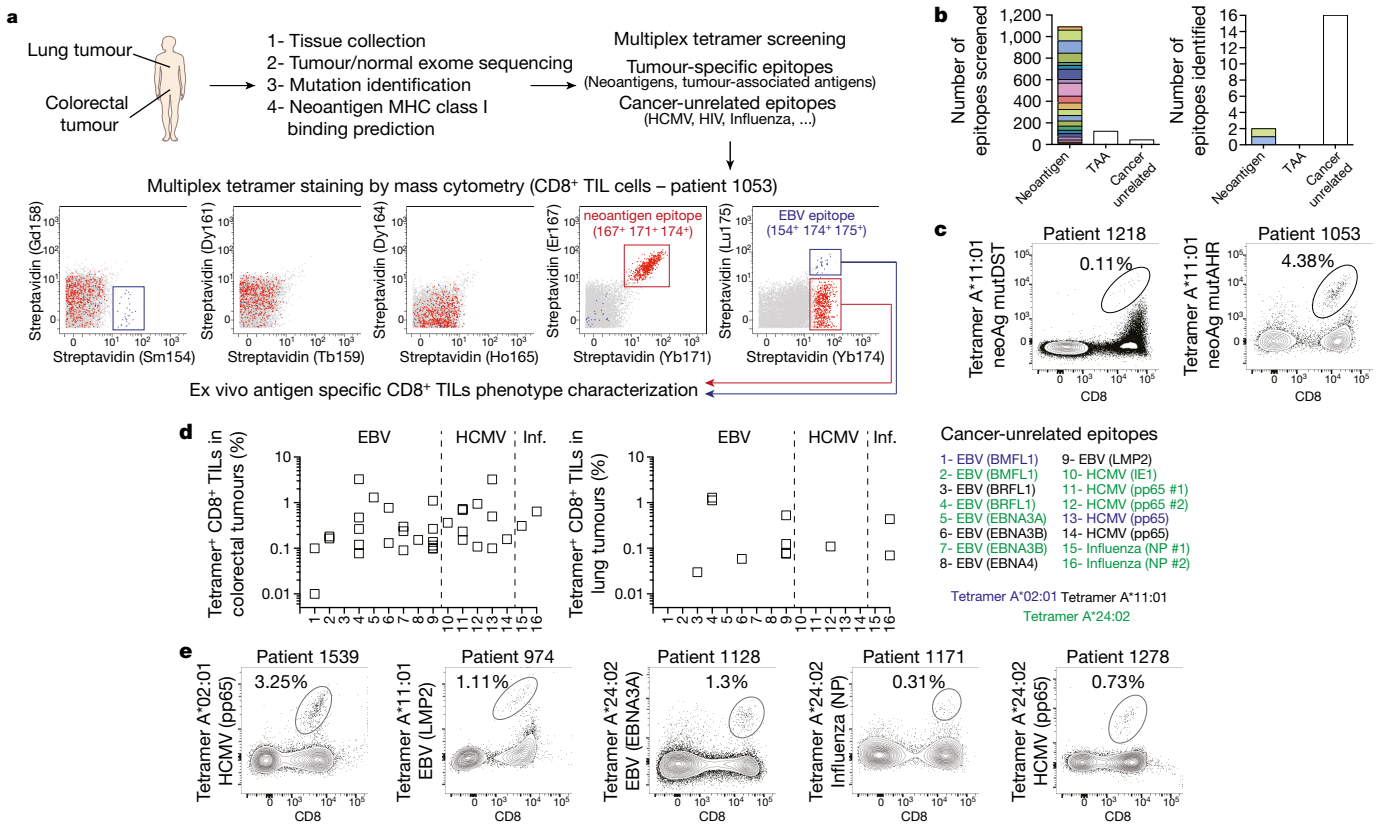


Fig. 2 | Tumour-specific and cancer-unrelated CD8$^+$ T cells infiltrate tumour tissues. a, Schematic for screening of neoantigens (NeoAg), tumour-associated antigens (TAA) and cancer-unrelated epitopes by mass cytometry coupled to multiplex MHC-tetramer staining. See also Extended Data Fig. 2 and Supplementary Tables 1–3 for examples and list of peptides. **b**, Total number of different MHC class I tetramers screened for neoantigens, TAA and cancer-unrelated epitopes by mass cytometry (left). Total number of different MHC class I tetramers identified for neoantigens, TAA and cancer-unrelated epitopes by mass cytometry (right). Neoantigens are colour-coded by patient. See also Supplementary Table 4 for patient information. **c**, Flow cytometry dot plots representing MHC tetramer$^+$ CD8$^+$ TILs identified from colorectal tumour CD3$^+$ TILs using mass cytometry screening. MHC-tetramer mutated (mut) DST epitope (mutated residue is underlined), ISDEMFKTFK; MHC-tetramer mutAHR epitope, GISQELPYK. Percentages represent the MHC tetramer$^+$ cells among CD8$^+$ TILs for each patient. Data from two independent experiments. **d**, Frequencies of cancer-unrelated CD8$^+$ TILs identified by mass cytometry and multiplex MHC-tetramer staining in lung tumours (right, $n = 11$ MHC tetramer$^+$ populations) or colorectal tumours (left, $n = 35$ MHC tetramer$^+$ populations). Peptide sequences are listed in Supplementary Table 3. Inf., influenza virus. **e**, Representative flow cytometry dot plot showing cancer-unrelated CD8$^+$ TILs specific for different epitopes, identified from colorectal tumour CD3$^+$ TILs using mass cytometry screening. Percentages are of MHC tetramer$^+$ cells among CD8$^+$ TILs for each patient. See also Extended Data Fig. 4. Data are from two independent experiments.

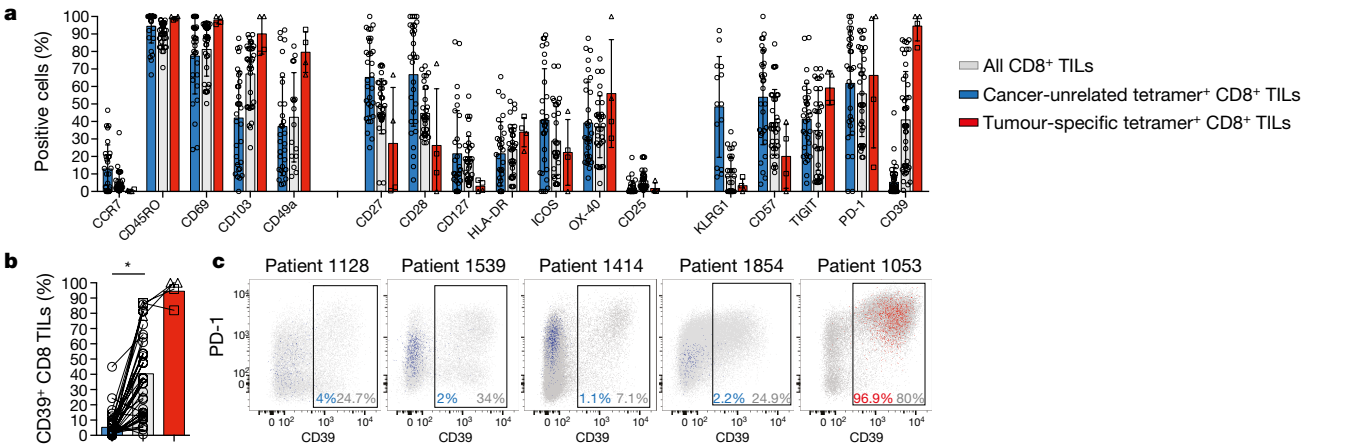


Fig. 3 | Cancer-unrelated CD8$^+$ TILs do not express CD39. a, Expression of markers by cancer-unrelated (blue, $n = 46$ biologically independent MHC tetramer$^+$ cells) and tumour-specific CD8$^+$ TILs (red, $n = 4$ biologically independent MHC tetramer$^+$ cells) in human tumours. Triangles represent neoantigen-specific CD8$^+$ TILs and squares represent tumour-specific CD8$^+$ TILs derived from an LELC (See Extended Data Figs. 3, 5). Data are mean ± s.d. from at least ten independent mass cytometry experiments. Each data point represents an antigen-specific population. **b**, Expression of CD39 by cancer-unrelated (blue, $n = 46$ biologically independent MHC tetramer$^+$ cells) or tumour-specific CD8$^+$ T cells (red, $n = 4$ biologically independent MHC tetramer$^+$ cells) with paired total CD8$^+$ TILs. Triangles represent neoantigen-specific CD8$^+$ TILs and squares represent tumour-specific CD8$^+$ TILs in an LELC (See Extended Data Fig. 3). Data are mean ± s.d.; paired two-tailed t-test; lung tumour only, $P = 0.0064$; colorectal tumour only, $^*P < 0.0001$. **c**, Representative flow cytometry dot plot representing the expression of PD-1 and CD39 by cancer-unrelated (blue) or tumour-specific CD8$^+$ TILs (red) identified from colorectal tumours by mass cytometry. Frequencies of CD39$^+$ cells among cancer-unrelated (blue), tumour-specific (red) or all CD8$^+$ TILs (grey) for each patient. Data are from two independent experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

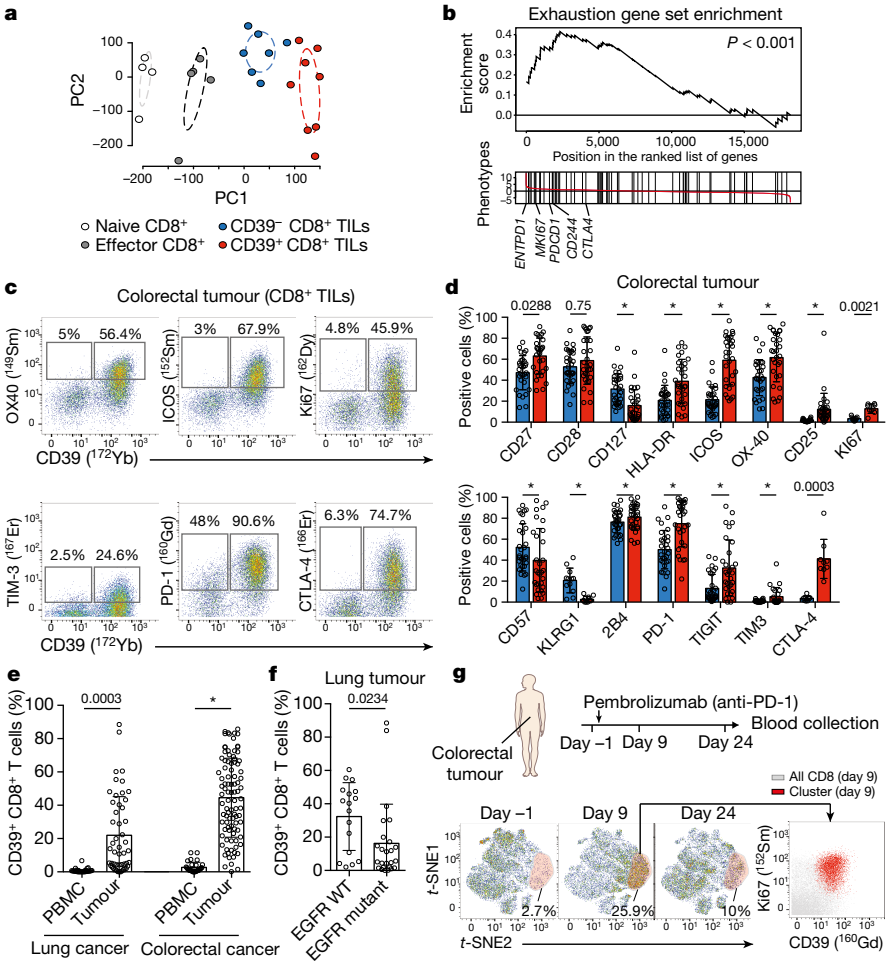


Fig. 4 | Comparative analysis of CD8$^+$ TILs stratified by CD39 expression. a, Projection of the whole transcriptome of sorted blood naive (CCR7$^+$ CD45RO$^-$, white; $n = 4$ patients), blood effector (CCR7$^-$ CD45RO$^+$, grey; $n = 5$ patients), tumour CD39$^-$ (blue, $n = 7$ patients) and CD39$^+$ CD8$^+$ TILs (red, $n = 8$ patients) using PCA. Ellipses represent the standard deviation around the centroid of a phenotype. See also Supplementary Table 5. **b**, Enrichment of the gene set for exhausted T cells[28,29] in CD39$^+$ CD8$^+$ TILs. Genes towards the left are enriched in CD39$^+$ CD8$^+$ TILs ($n = 8$ patients), genes towards the right are enriched in CD39$^-$ CD8$^+$ TILs ($n = 7$ patients). Two-sided GSEA empirical test. See also Extended Data Figs. 6, 7. **c**, Mass cytometry dot plots representing expression of OX-40, ICOS, Ki67, TIM-3, PD-1 and CTLA-4 according to CD39 status by CD8$^+$ TILs. Representative data from one patient with a colorectal tumour. Data are from at least ten independent mass cytometry experiments. **d**, Frequencies of the expression of activation markers (top) and inhibitory markers (bottom) by CD39$^-$ (blue) and CD39$^+$ (red) CD8$^+$ TILs in colorectal tumours ($n = 9$ to 45 patients). Data are mean ± s.d. from at least ten independent mass cytometry experiments. Two-tailed paired *t*-test. *$P < 0.0001$. See also Extended Data Fig. 8. **e**, Frequencies of CD39$^+$ CD8$^+$ T cells in lung tumours ($n = 50$ patients), colorectal tumours ($n = 94$ patients) and the matched PBMCs. Data are mean ± s.d. **f**, Frequencies of CD39$^+$ CD8$^+$ T cells in lung tumours stratified by mutation status of EGFR. Wild-type (WT) EGFR, $n = 17$ patients; mutant EGFR, $n = 25$ patients. Data are mean ± s.d., two-tailed unpaired *t*-test. **g**, Schematic for blood collection time point in colorectal cancer patients treated with pembrolizumab (anti-PD-1 treatment) (top). *t*-SNE map of CD8$^+$ T cells isolated from PBMC at −1, 9 and 24 days relative to pembrolizumab treatment (bottom).

In our study cohort, the frequency of CD39$^+$ CD8$^+$ TIL cells was significantly higher in patients with EGFR-wild-type tumours (32.3% ± 20.35%) in comparison to those with EGFR-mutant tumours (16.24% ± 23.51%), where these cells were virtually absent (<2% of CD8$^+$ TIL cells) in 50% of these patients (Fig. 4f). These data suggest that the poor observed response to anti-PD-1 treatment could be associated with the abundance of bystander CD39$^-$ CD8$^+$ TIL cells in patients with EGFR-mutant tumours.

We also obtained peripheral blood from a patient with microsatellite-instable metastatic colorectal cancer who realized a rapid clinical and radiological response with anti-PD-1 treatment (pembrolizumab). The proliferation marker Ki67 enables identification and monitoring of the immunological response during anti-PD-1 treatment in peripheral blood[33]. Of note, the proliferating CD8$^+$ T cells in the blood from this patient were characterized by high expression of CD39 (Fig. 4g). These data are consistent with an expansion of a CD39$^+$ population in the peripheral blood of a patient responding to anti-PD1 treatment. This suggests that serial changes in CD39$^+$ T cells may be an early blood-based readout of anti-tumour-specific CD8$^+$ responses and may indicate a promising role for CD39 in monitoring immune checkpoint therapy.

In summary, our results demonstrate that human CD8$^+$ TILs can not only be specific for tumour antigens but may also contain a population that recognizes cancer-unrelated epitopes, characterized by an absence of CD39 expression. Our data show that tumours can possess a wide range of frequencies and phenotypes of bystander infiltrating CD8$^+$ T cells. Conversely, although partially based on very limited examples of tumour-specific T cell populations (two neoantigen-specific populations from two patients and two tumour-specific populations from an uncommon LELC tumour), our data also suggest that CD39 could be useful as a marker of tumour-specific CD8$^+$ T cells, which could be exploited for the development of novel biomarkers or therapeutics. Future work will determine the extent to which CD39 is a useful marker of tumour specificity.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0130-2.

Received: 19 September 2017; Accepted: 18 April 2018;
Published online: 16 May 2018

1. Rosenberg, S. A. & Restifo, N. P. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* **348**, 62–68 (2015).
2. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).
3. Ott, P. A. et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).
4. Sahin, U. et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547**, 222–226 (2017).
5. Lavin, Y. et al. Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses. *Cell* **169**, 750–765.e717 (2017).
6. Chevrier, S. et al. An immune atlas of clear cell renal cell carcinoma. *Cell* **169**, 736–749.e718 (2017).
7. Gubin, M. M. et al. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515**, 577–581 (2014).
8. Fehlings, M. et al. Checkpoint blockade immunotherapy reshapes the high-dimensional phenotypic heterogeneity of murine intratumoural neoantigen-specific CD8$^+$ T cells. *Nat. Commun.* **8**, 562 (2017).
9. Liu, X. S. & Mardis, E. R. Applications of immunogenomics to cancer. *Cell* **168**, 600–612 (2017).
10. Yarchoan, M., Johnson, B. A. III, Lutz, E. R., Laheru, D. A. & Jaffee, E. M. Targeting neoantigens to augment antitumour immunity. *Nat. Rev. Cancer* **17**, 209–222 (2017).
11. Zheng, C. et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* **169**, 1342–1356.e1316 (2017).
12. Newell, E. W. et al. Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization. *Nat. Biotechnol.* **31**, 623–629 (2013).
13. Khodadoust, M. S. et al. Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature* **543**, 723–727 (2017).
14. Cohen, C. J. et al. Isolation of neoantigen-specific T cells from tumor and peripheral lymphocytes. *J. Clin. Invest.* **125**, 3981–3991 (2015).
15. McGranahan, N. et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).
16. Chheda, Z. S. et al. Novel and shared neoantigen derived from histone 3 variant H3.3K27M mutation for glioma T cell therapy. *J. Exp. Med.* **215**, 141–157 (2018).
17. Flecken, T. et al. Immunodominance and functional alterations of tumor-associated antigen-specific CD8$^+$ T-cell responses in hepatocellular carcinoma. *Hepatology* **59**, 1415–1426 (2014).
18. Koido, S. et al. Induction of antigen-specific CD4$^-$ and CD8-mediated T-cell responses by fusions of autologous dendritic cells and metastatic colorectal cancer cells. *Int. J. Cancer* **117**, 587–595 (2005).
19. Liu, G. et al. HER-2, gp100, and MAGE-1 are expressed in human glioblastoma and recognized by cytotoxic T cells. *Cancer Res.* **64**, 4980–4986 (2004).
20. Matsuzaki, J. et al. Tumor-infiltrating NY-ESO-1-specific CD8$^+$ T cells are negatively regulated by LAG-3 and PD-1 in human ovarian cancer. *Proc. Natl Acad. Sci. USA* **107**, 7875–7880 (2010).
21. Ahmadzadeh, M. et al. Tumor antigen-specific CD8 T cells infiltrating the tumor express high levels of PD-1 and are functionally impaired. *Blood* **114**, 1537–1544 (2009).
22. Johnston, R. J. et al. The immunoreceptor TIGIT regulates antitumor and antiviral CD8$^+$ T cell effector function. *Cancer Cell* **26**, 923–937 (2014).
23. Gros, A. et al. PD-1 identifies the patient-specific CD8$^+$ tumor-reactive repertoire infiltrating human tumors. *J. Clin. Invest.* **124**, 2246–2259 (2014).
24. Erkes, D. A. et al. Virus-specific CD8$^+$ T cells infiltrate melanoma lesions and retain function independently of PD-1 expression. *J. Immunol.* **198**, 2979–2988 (2017).
25. Canale, F. P. et al. CD39 expression defines cell exhaustion in tumor-infiltrating CD8$^+$ T cells. *Cancer Res.* **78**, 115–128 (2018).
26. Gupta, P. K. et al. CD39 expression identifies terminally exhausted CD8$^+$ T cells. *PLoS Pathog.* **11**, e1005177 (2015).
27. Soares, A. et al. Novel application of Ki67 to quantify antigen-specific in vitro lymphoproliferation. *J. Immunol. Methods* **362**, 43–50 (2010).
28. Wherry, E. J. et al. Molecular signature of CD8$^+$ T cell exhaustion during chronic viral infection. *Immunity* **27**, 670–684 (2007).
29. Baitsch, L. et al. Exhaustion of tumor-specific CD8$^+$ T cells in metastases from melanoma patients. *J. Clin. Invest.* **121**, 2350–2360 (2011).
30. Pauken, K. E. & Wherry, E. J. Overcoming T cell exhaustion in infection and cancer. *Trends Immunol.* **36**, 265–276 (2015).
31. Nahar, R. et al. Elucidating the genomic architecture of Asian EGFR-mutant lung adenocarcinoma through multi-region exome sequencing. *Nat. Commun.* **9**, 216 (2018).
32. Gainor, J. F. et al. EGFR mutations and ALK rearrangements are associated with low response rates to PD-1 pathway blockade in non-small cell lung cancer: a retrospective analysis. *Clin. Cancer Res.* **22**, 4585–4593 (2018).
33. Huang, A. C. et al. T-cell invigoration to tumour burden ratio associated with anti-PD-1 response. *Nature* **545**, 60–65 (2017).

Acknowledgements The authors thank all members of the Newell laboratory, the SIgN community, the flow cytometry platform, all participating patients, the clinical research coordinators from NCCS, S. Quezada and J. Reading. This study was funded by A-STAR/SIgN core funding and A-STAR/SIgN immunomonitoring platform funding assigned to E.W.N. as well as NMRC/CSA-INV/0001/2014 grant (GIS) and core A-STAR/GIS funding to I.B.T. and the lung TCR grant NMRC/TCR/007-NCC/2013.

Reviewer information *Nature* thanks N. Haining, J. Wherry and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.S. designed and performed research, analysed data and wrote the paper. E.B. analysed transcriptomic and clinical data and reviewed the paper. M.F. provided technical help with CyTOF, performed experiments and reviewed the paper. C.Y.L., K.W.W.T., J.P.S.Y. and N.A. helped to process samples and perform experiments. S.-L.K., E.T., C.F., R.M., M.T., W.T.L., C.K.T., B.-H.O., T.K., A.T., T.K.H.L., E.H.T., D.S.W.T., I.B.T., H.K., K.D., A.L., M.P., Y.Y.L. provided samples and discussed data. R.N., T.Z., A.J.K., A.M.H. and W.Z. performed neoantigen predictions. E.W.N. initiated and led the project, developed scripts for CyTOF analysis and wrote the paper.

Competing interests E.W.N. is a board director and shareholder of immunoSCAPE. M.F. is director, scientific affairs and shareholder of immunoSCAPE. All other authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0130-2
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0130-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Y.S. or E.W.N.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Human samples. PBMCs, adjacent normal and tumour samples (lung and colon) were obtained from patients with colorectal cancer or lung cancer (See Supplementary Table 6). The use of human tissues was approved by the appropriate institutional research boards, A*STAR and the Singapore Immunology Network.

Cell isolation. Samples were prepared as previously described[34]. In brief, tissues were mechanically dissociated into small pieces and incubated at 37 °C for 15 to 40 min in DMEM + collagenase IV (1 mg/ml) + DNase (15 μg/ml). Digestion was stopped by addition of RPMI containing 5% FBS. Dissociated tissues were filtered and washed in RPMI 5% + DNase (15 μg/ml) FBS. All samples were cryopreserved in 90% FBS + 10% DMSO and stored in liquid nitrogen.

MHC monomer production and heavy metal labelling of streptavidin and antibodies. Purified antibodies lacking carrier proteins were purchased according to the list in Supplementary Table 7. Antibody conjugation was performed according to the protocol provided by Fluidigm. Streptavidin was labelled as previously described[12]. HLA-A*11:01 and HLA-A*24:02 monomer was refolded with appropriate UV-cleavable peptide and biotinylated as described[35,36].

Tumour-associated antigen and neoantigen epitopes prediction. Binding of epitopes of common TAA protein[37] to HLA-A*11:01 or HLA-A*24:02 was predicted using the NetMHC3.4 algorithm[38]. Candidate epitopes (8 to 11 amino acids) with an affinity <500 nM were selected. For neoantigens, exome and RNA sequencing were performed at Genome institute of Singapore (GIS). Tumour and adjacent normal tissues were harvested by pathologists following surgical resection. DNA and RNA were extracted from clinical specimens using the Qiagen Allprep DNA/RNA kit. After the QC step, DNA was sonicated to shorter fragments using the Covaris system and end-repaired and ligated with sequencing adapters. Agilent SureSelect Human All Exon V6 was used for exome capture before sequencing. The Illumina Hiseq 4000 platform was used for whole exome sequencing (WES). For DNA sequencing, short sequence reads were mapped against human reference genome GRCh37 using BWA-MEM with default parameters[39]. Following GATK best practice[40], PCR duplicates were first removed and subsequently realigned and recalibrated (available at: https://github.com/gis-rpd/pipelines, GATK v.3.5). Somatic mutations (SNVs) were identified using MuTect (version 1.1.7)[41]. Somatic insertions and deletions (indels) were called using Strelka with default parameters[42]. The raw RNA-seq data was mapped to the reference genome using STAR. The data were ported to RSEM to generate count data and subsequently normalized using DESeq2. Non-synonymous mutations and indels were used to predict the mutated protein sequences using ANNOVAR[43]. A list of peptides 8–11 amino acids in length that are not seen in normal cells (covering either the mutated position or the novel C-terminal sequence due to frame-shift indels) was generated. Subsequently, the binding affinity of every mutant peptide and its corresponding wild-type peptide to the predicted HLA-A alleles was estimated using netMHC3.4. The putative neo-antigens were identified as the mutant peptide with a predicted binding strength of <500 nM and lower binding strength than that of its corresponding wild type. All peptides (predicted TAA, published TAA, neoantigens) were synthetized at Mimotopes and diluted in DMSO to the final concentration of 10 mM.

mRNA sequencing data analysis. The paired-end RNA-seq reads from HiSeq 4000 were mapped to the human GRCh38/HG19 reference genome using the STAR software tool. The mapped paired-end reads were summarized to gene level using featureCounts v.1.5.0-p1 software tool[44] and with GENCODE v26 gene annotation[45]. Genes with read count less than five in fewer than two samples in all cell populations were filtered out from further analysis. Limma-voom[46] pipeline was used for differentially expressed gene (DEG) analysis. DEGs from comparisons between different cell populations were selected with Benjamini–Hochberg adjusted P value of <0.05. All analyses were done in R v.3.1.2[47]. We ran PCA using the prcomp function from the stats (v.3.4.2) R package, using the whole transcriptomic data. Data ellipses were drawn using the ordiellipse function of the *vegan* (v.2.4-5) R package. We used the HTSanalyzeR package (v.2.26.0) to run GSEA and hypergeometric tests on gene collections from the Gene Ontology Biological Processes database, filtered for gene sets with at least 20 genes present in our dataset. For GSEA we used 1000 permutations to estimate P values and applied corrections for multiple tests using the Benjamini–Hochberg procedure. Results were displayed by plotting the 40 most significant pathways using the enrichment map package.

We performed hierarchical clustering on the Pearson correlation matrix of the genes whose variance was in the top 25%, using Ward's linkage criterion and $1-r$ as the distance function, where r is a pairwise Pearson's correlation coefficient. The resulting dendrogram was cut into 10 subtrees to obtain gene modules.

To compute clonality indices, we first converted the measured counts for TCR to frequencies. We then computed Shannon's entropy normalized for the number of unique pairings n by $H = \sum_{i=1}^{n} -\log(f_i) \times f_i / \log(n)$, where f_i is the frequency of the pairing n, as previously proposed[48], H varies from 1 for a uniform distribution to 0 for an entirely clonal distribution. Clonality C was thus defined as $C = 1 - H$ and varies from 0 to 1, with 1 indicating high clonality.

Multiplexing tetramer preparation and staining. For multiplex MHC-tetramer staining, each tetramer was labelled with a combination of three metal-labelled streptavidins. Using ten different metal-labelled streptavidins, 120 possible combinations (10 choose 3) were generated. Each specific combination was associated with a different peptide. Each metal-labelled streptavidin (50 μg/ml, 20 μl) was mixed for each combination using an automated pipetting device (TECAN). In a 96 well plate, 5 μl peptide (1 mM) was add to 100 μl HLA monomer (100 μg/ml, diluted in PBS), with a different peptide in each well. The plate was exposed to UV (365 nm) for 10 min for peptide exchange and left overnight at 4 °C. For the tetramerisation, each peptide–MHC complex–metal-labelled streptavidin combination (50 μg/ml) was added in three steps (3 × 10 μl) according to the coding scheme. Then, tetramerized peptide-MHC complexes were incubated with free biotin for 10 min (10 μM) (Sigma). All different tetramers were combined and concentrated using a 50-kDa Amicon filter (Millipore) to a final volume of 500 μl. Then, 500 μl PBS, 1% BSA, 0.02% sodium azide was added. Before staining, the tetramer cocktail was filtered using a 0.1-μm filter (Millipore).

Frozen samples were thawed and washed in RPMI, 10% FBS, 15 μg/ml DNase. Cells were incubated for 1 h at room temperature with the tetramer cocktail. Cells were then stained with 5 μM cisplatin (viability marker) in PBS for 5 min as previously described[34]. Cells were then incubated with antibody cocktail for 15 min (See Supplementary Table 5 for clone list and metals) and fixed in 2%PFA.

Data analysis and *t*-SNE. After mass cytometry (CyTOF) acquisition, which was performed as previously described[49,50], any zero values were randomized using a uniform distribution of values between zero and −1 using an R script (as was the default operation of previous CyTOF software). Note also that all other integer values measured by the mass cytometer are randomized in a similar fashion by default. The signal of each parameter was then normalized based on the EQ beads (Fluidigm) as previously described[51]. Cells were manually de-barcoded using FlowJo (Tree Star). Samples were then used for *t*-SNE analysis similar to that previously described[49,50] using custom R scripts based on the 'flowCore' and 'Rtsne' (using CRAN R packages that perform the Barnes–Hut implementation of *t*-SNE)[52,53]. In R, all data were transformed using the logicleTransform function (flowCore package) using parameters: $w = 0.25$, $t = 16409$, $m = 4.5$, $a = 0$ to roughly match scaling historically used in FlowJo. For heatmaps, median intensity corresponds to a logical data scale using formula previously described[54]. The colours in the heat map represent the measured means intensity value of a given marker in a given cluster. A four-colour scale is used with black–blue indicating low expression values, green–yellow indicating intermediately expressed markers, and red representing highly expressed markers.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

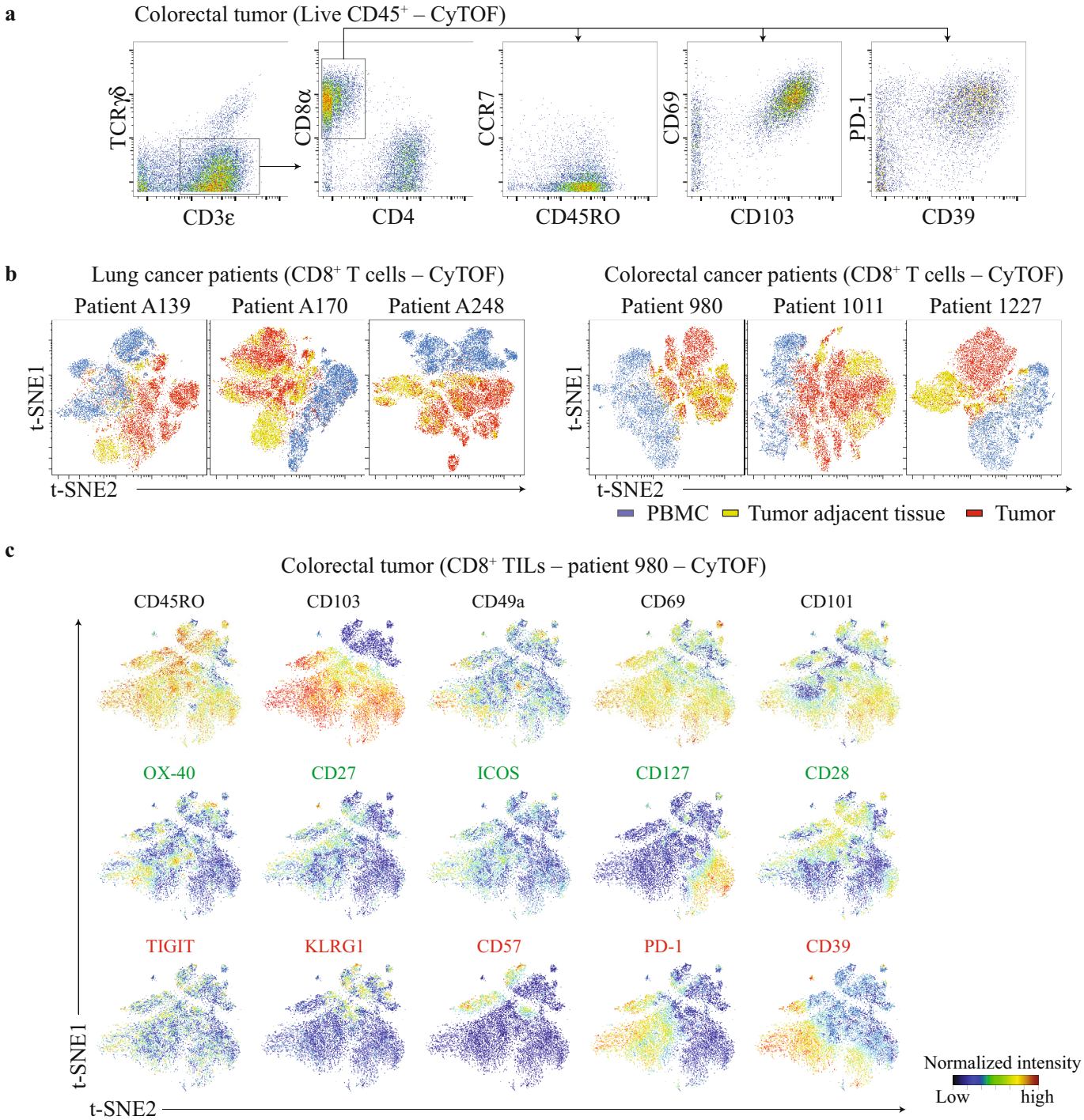
Data availability. Sequence data that support the findings of this study have been deposited in the Gene Expression Omnibus under accession number GSE113590. Mass cytometry data have been deposited in FlowRepository (https://flowrepository.org) under accession link FR-FCM-ZYWM. Further data that support the findings of this study are available from the corresponding authors upon reasonable request.

34. Simoni, Y. et al. Human innate lymphoid cell subsets possess tissue-type based heterogeneity in phenotype and frequency. *Immunity* **46**, 148–161 (2017).
35. Bakker, A. H. et al. Conditional MHC class I ligands and peptide exchange technology for the human MHC gene products HLA-A1, -A3, -A11, and -B7. *Proc. Natl Acad. Sci. USA* **105**, 3825–3830 (2008).
36. Toebes, M. et al. Design and use of conditional MHC class I ligands. *Nat. Med.* **12**, 246–251 (2006).
37. Coulie, P. G., Van den Eynde, B. J., van der Bruggen, P. & Boon, T. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat. Rev. Cancer* **14**, 135–146 (2014).
38. Lundegaard, C. et al. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res.* **36**, W509-512 (2008).
39. Li, Y. & Ghosh, D. A two-step hierarchical hypothesis set testing framework, with applications to gene expression data on ordered categories. *BMC Bioinformatics* **15**, 108 (2014).
40. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
41. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
42. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
43. Wang, S. et al. Label-free imaging, detection, and mass measurement of single viruses by surface plasmon resonance. *Proc. Natl Acad. Sci. USA* **107**, 16028–16032 (2010).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

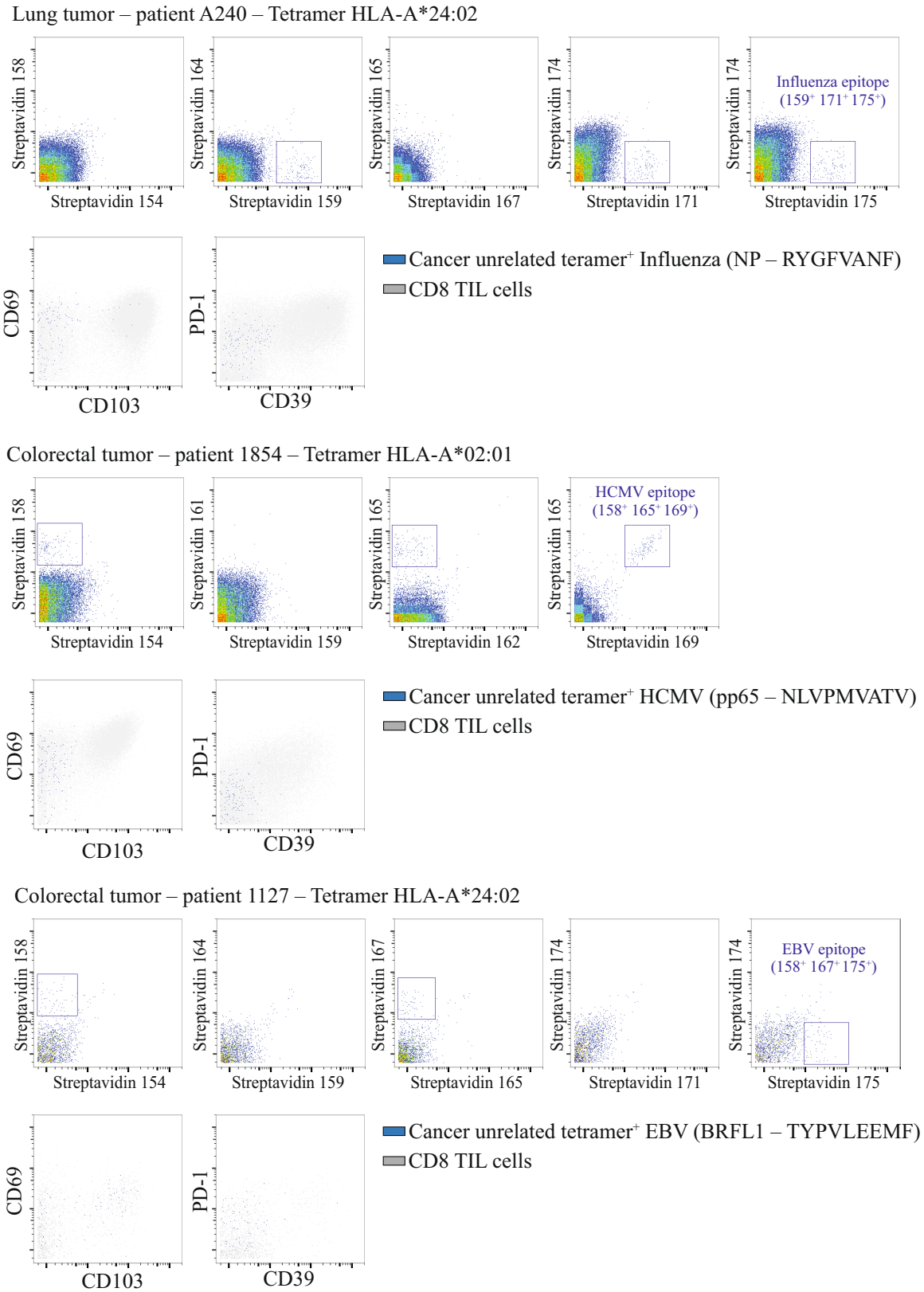
44. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
45. Harrow, J., et al GENCODE: producing a reference annotation for ENCODE. *Genome biology* **7**, 1–9 (2006).
46. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
47. R Core Team. *R: A Language and Environment for Statistical Computing* https://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2013).
48. Nunes-Alves, C. et al. Human and murine clonal $CD8^+$ T cell expansions arise during tuberculosis because of TCR selection. *PLoS Pathog.* **11**, e1004849 (2015).
49. Becher, B. et al. High-dimensional analysis of the murine myeloid cell system. *Nat. Immunol.* **15**, 1181–1189 (2014).
50. Wong, M. T. et al. Mapping the diversity of follicular helper T cells in human blood and tonsils using high-dimensional mass cytometry analysis. *Cell Reports* **11**, 1822–1833 (2015).
51. Finck, R. et al. Normalization of mass cytometry data with bead standards. *Cytometry A* **83**, 483–494 (2013).
52. Amir, A. D. et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
53. Van der Maaten, L. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 85 (2008).
54. Moore, W. A. & Parks, D. R. Update for the logicle data scale including operational code implementations. *Cytometry A* **81**, 273–277 (2012).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



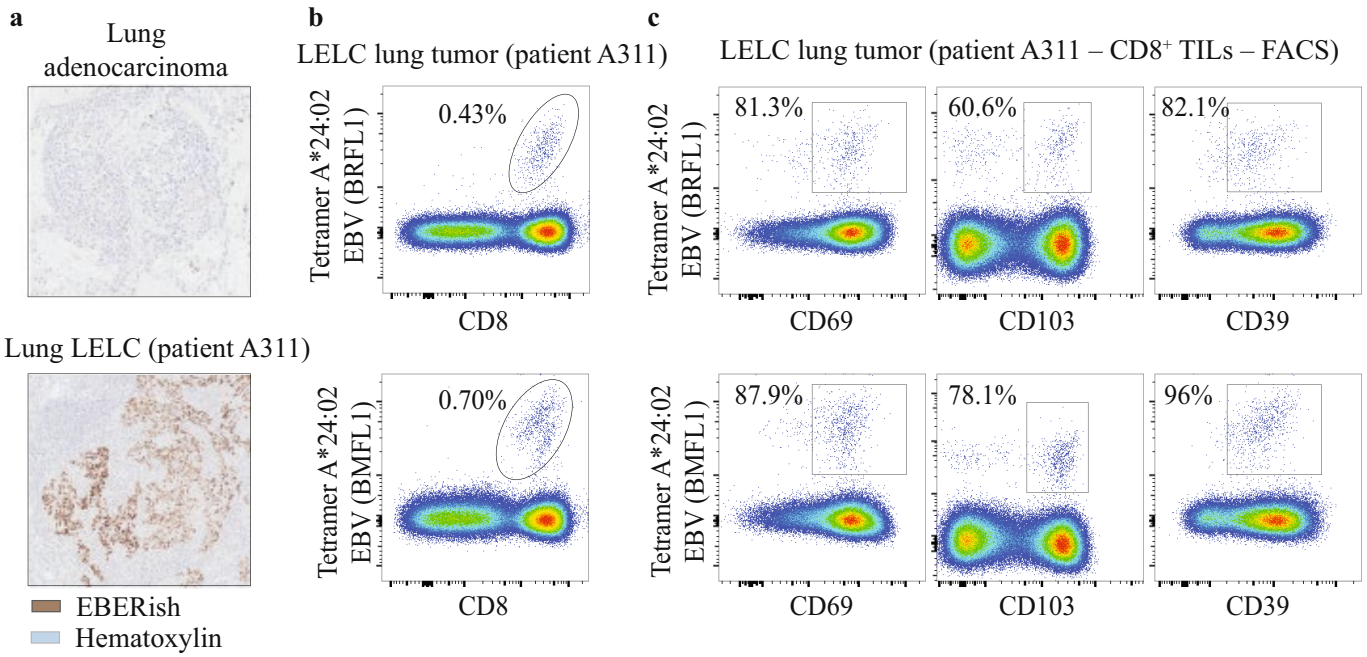
Extended Data Fig. 1 | Tumour-infiltrating CD8$^+$ T cells are heterogeneous. a, Mass cytometry dot plots showing expression of markers expressed by tumour-infiltrating CD8$^+$ TILs (gated on CD45$^+$ cisplatin$^-$). Data from at least 10 independent mass cytometry experiments. Representative data from one patient. **b**, For high-dimensional assessment of CD8$^+$ T cell heterogeneity, we used *t*- SNE, which accounts for nonlinear relationships between markers and projects high dimensional data into a low dimensional space by making a pairwise comparison of cellular phenotypes to optimally plot similar cells near each other (see Methods). Parallel analysis of CD8$^+$ T cells from PBMCs (blue), tumour-adjacent tissue (yellow) and tumour tissue (red) from lung or colorectal cancer patients was performed, which allows for accurate comparison of the phenotypes of cells from each of these sample types. *t*-SNE map of CD8$^+$ T cells isolated from PBMC (blue), tumour adjacent tissue (lung or colon, yellow) and tumour (red). *t*-SNE was performed separately on each patient. Data are from at least 10 independent mass cytometry experiments. Representative data from three patients for each malignancy. **c**, *t*-SNE analyses focusing only on CD8$^+$ T cells from tumour infiltrates to explore the heterogeneity of CD8$^+$ TILs within individual patient tumours. *t*-SNE map of CD8$^+$ TILs isolated from a colorectal tumour. Data are from at least 10 independent mass cytometry experiments. Representative data from one patient.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



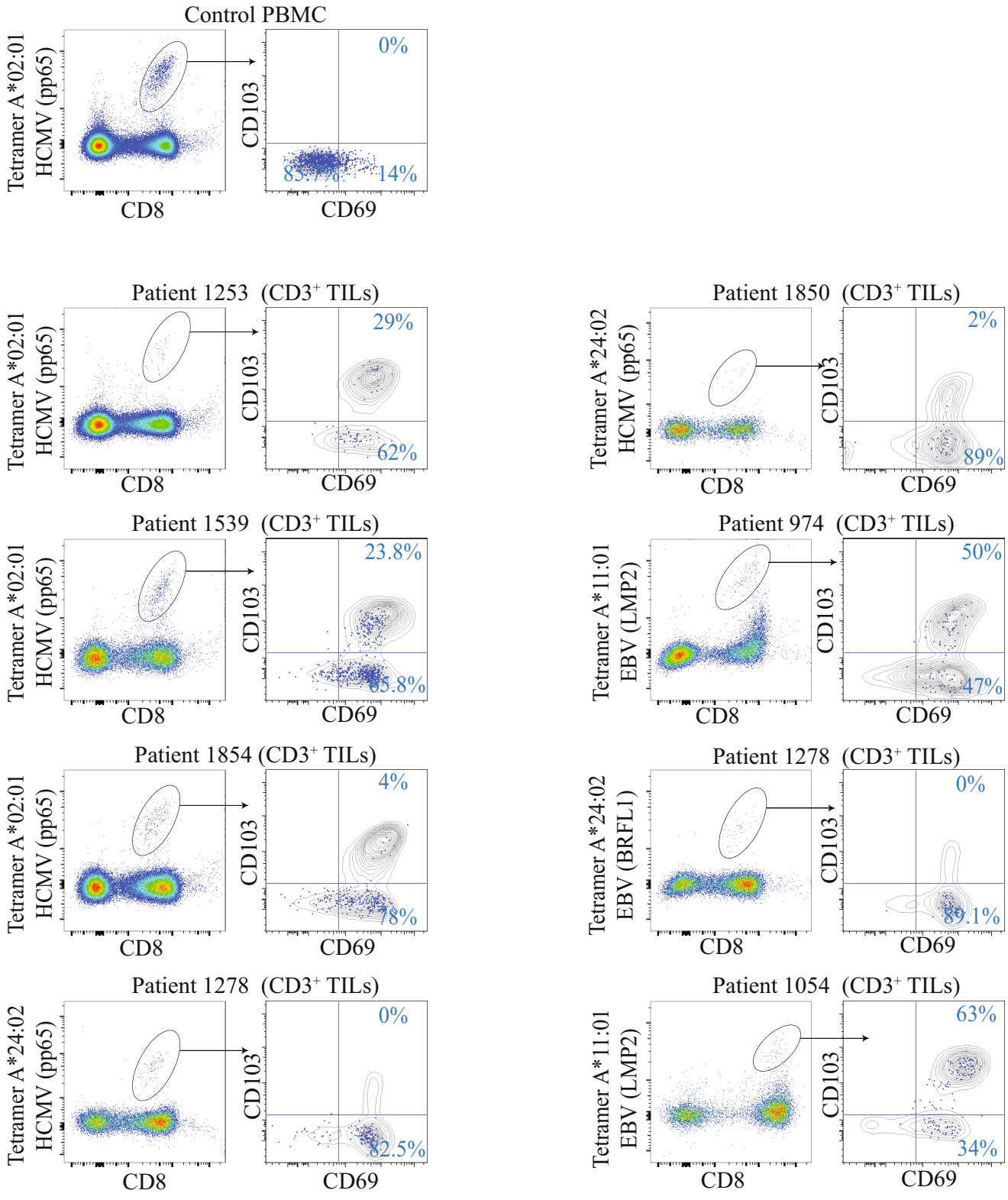
Extended Data Fig. 2 | Multiplex tetramer staining by mass cytometry. To investigate the antigen specificity of CD8$^+$ TILs, we performed multiplex MHC-tetramer staining as reported previously (see Methods). By using a three-metal coding scheme, we encoded up to 120 different tetramers specific for neoantigens, tumour-associated antigens (TAA) and cancer-unrelated epitopes (peptide list is shown in Supplementary Tables 1–3). Data are from at least 10 independent mass cytometry experiments (See Fig. 2). Representative data from three patients. Each MHC tetramer$^+$ cell population is positive for a unique code, composed of three differently-labelled streptavidin populations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



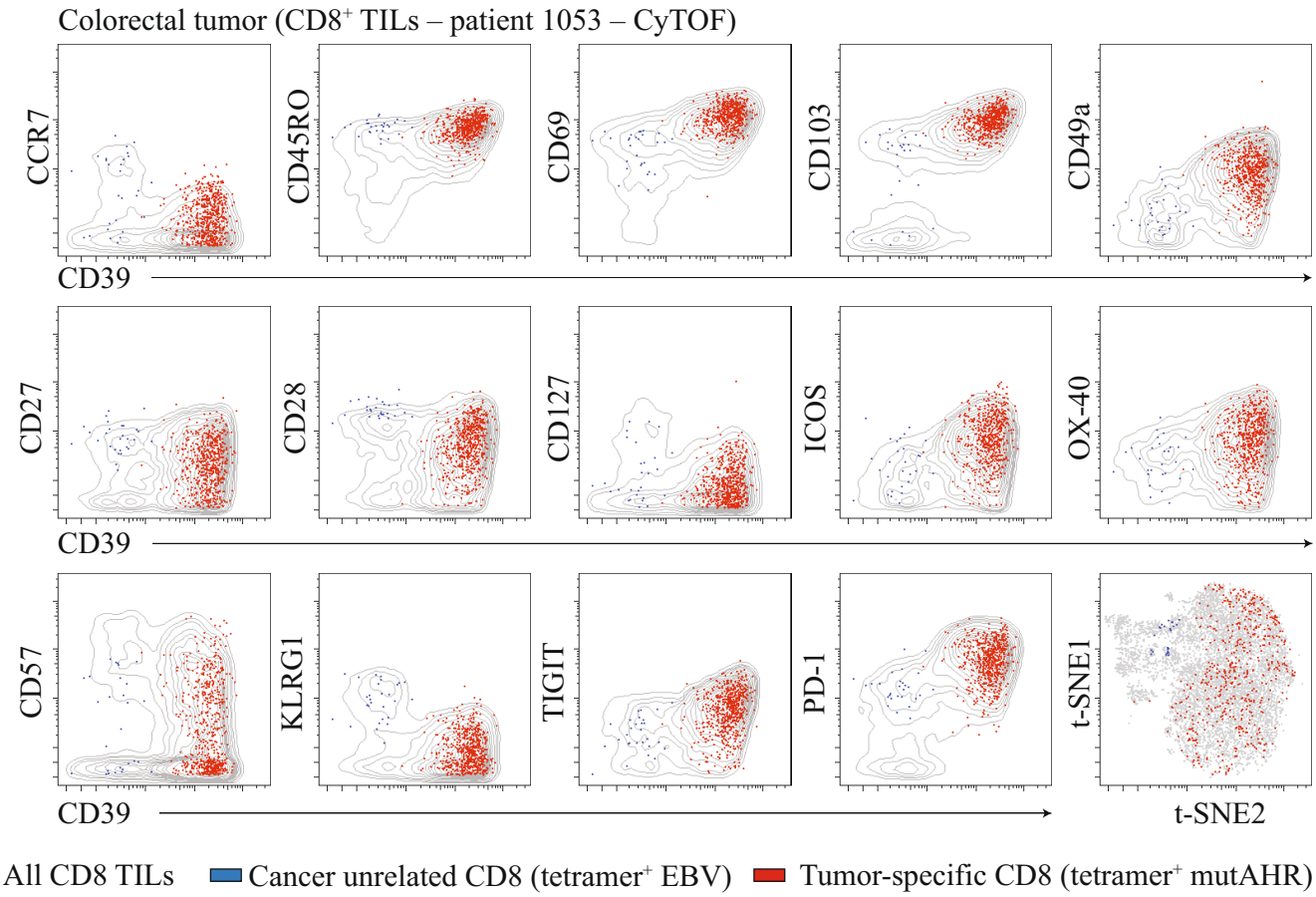
Extended Data Fig. 3 | Identification and phenotypic profiles of LELC tumour-specific CD8$^+$ TILs. a, Immunohistochemistry of lung adenocarcinoma and lung from LELC, stained with haematoxylin (blue) and EBV-encoded small RNA in situ hybridation (EBERish, brown). Primary LELC is rare and often associated with EBV infection in lung epithelial cells. Using EBERish staining on tissue sections, we confirmed the presence of EBV virus in tumour cells from patient A311. Data are from one experiment. **b**, Flow dot plot representing two populations specific for EBV-derived peptides (parts of the BRFL1 and BMFL1 proteins) infiltrating a LELC tumour. Since EBV proteins are presented by tumour cells, CD8$^+$ TILs specific for EBV epitope are also tumour specific. Frequency of MHC tetramer$^+$ cells among CD8$^+$ TILs. Data are from two independant experiments. Data from patient A311. **c**, Flow dot plot representing expression of CD69, CD103 and CD39 by LELC-specific CD8$^+$ TILs. As observed for neoantigen-specific CD8$^+$ TIL cells (Figs. 2, 3), LELC tumour-specific CD8$^+$ TILS express CD39. Frequency among MHC tetramer$^+$ cells. Data are from two independent experiments. Data from patient A311.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



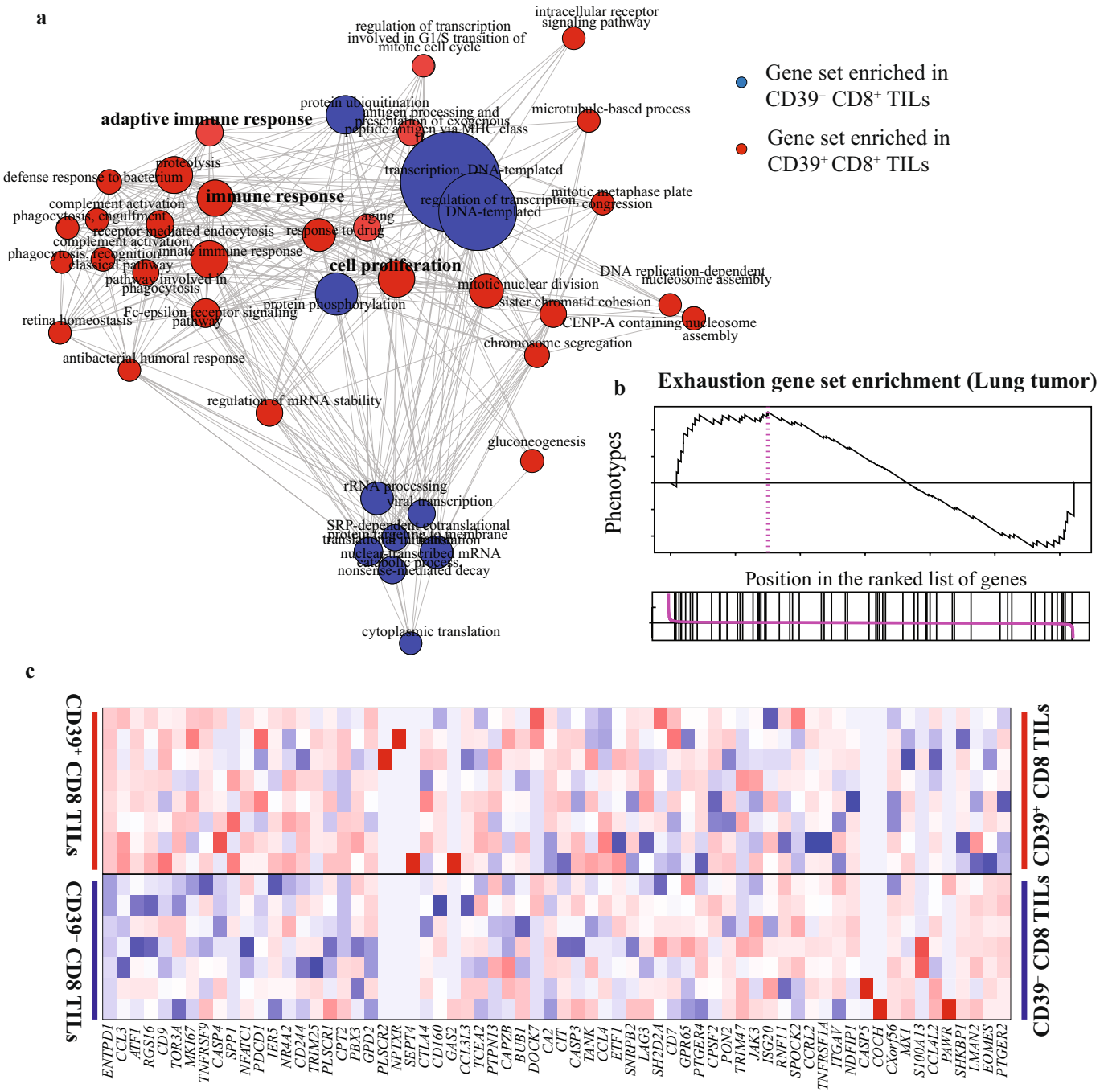
Extended Data Fig. 4 | Cancer-unrelated CD8$^+$ TILs express tissue-resident markers. Flow dot plot representing cancer-unrelated CD8$^+$ TILs specific for different viral epitopes, and expression of CD69 and CD103 by these cells identified by mass cytometry screening. Data are from three independent experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



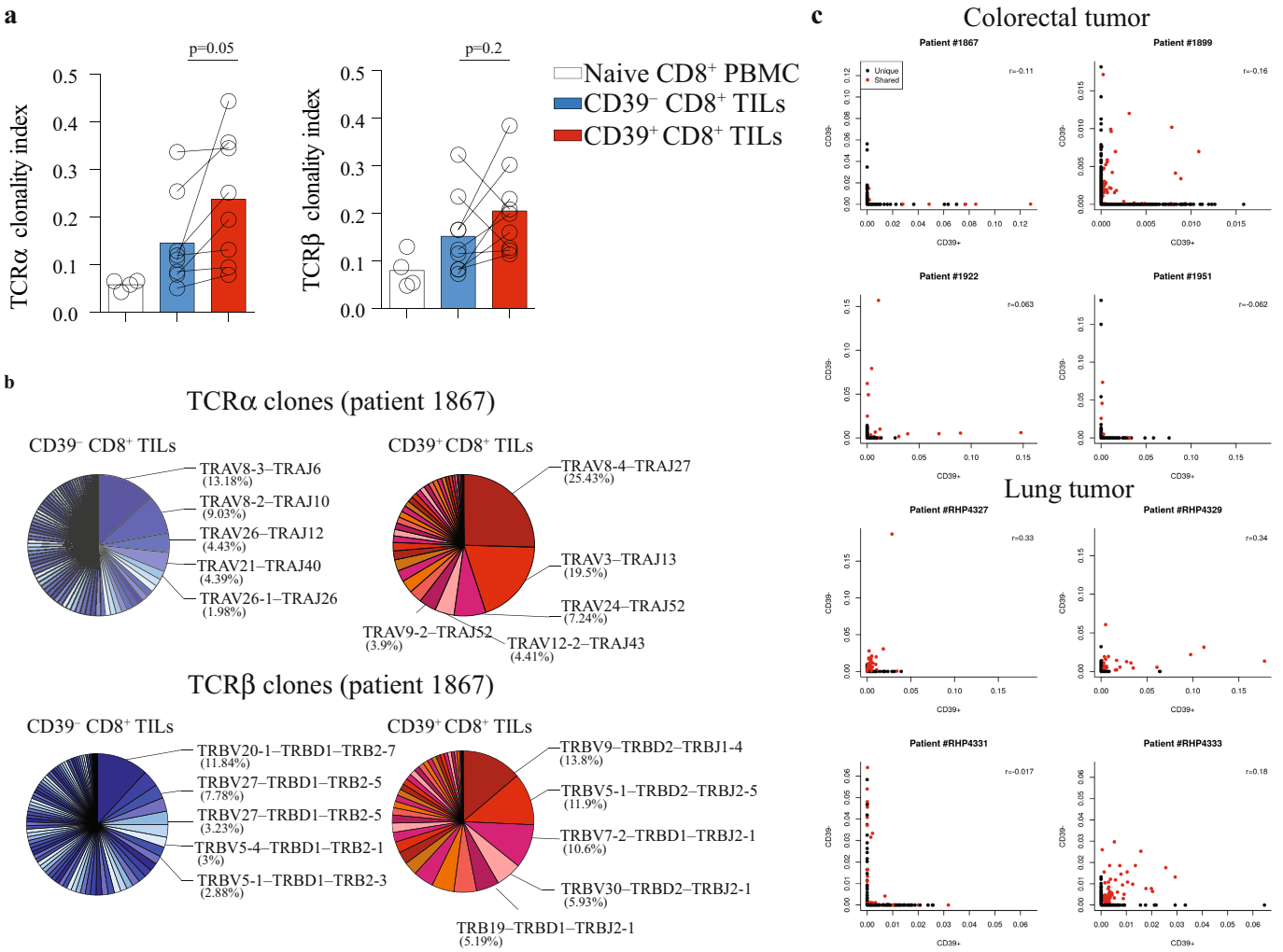
Extended Data Fig. 5 | Phenotypes of cancer-unrelated and tumour-specifc CD8+ TILs. Mass cytometry dot plot representing the expression of memory, co-activatory or inhibitory markers by cancer-unrelated $CD8^+$ TILs (blue, EBV-specific MHC $tetramer^+$, SSCSSCPLSK), neo-antigen specific $CD8^+$ TILs (red, mutAHR $tetramer^+$, GISQELPYK) and total $CD8^+$ TILs cells from colorectal tumour (See Fig. 2). Data are from at least ten independent mass cytometry experiments. Representative data from one individual, patient 1053.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



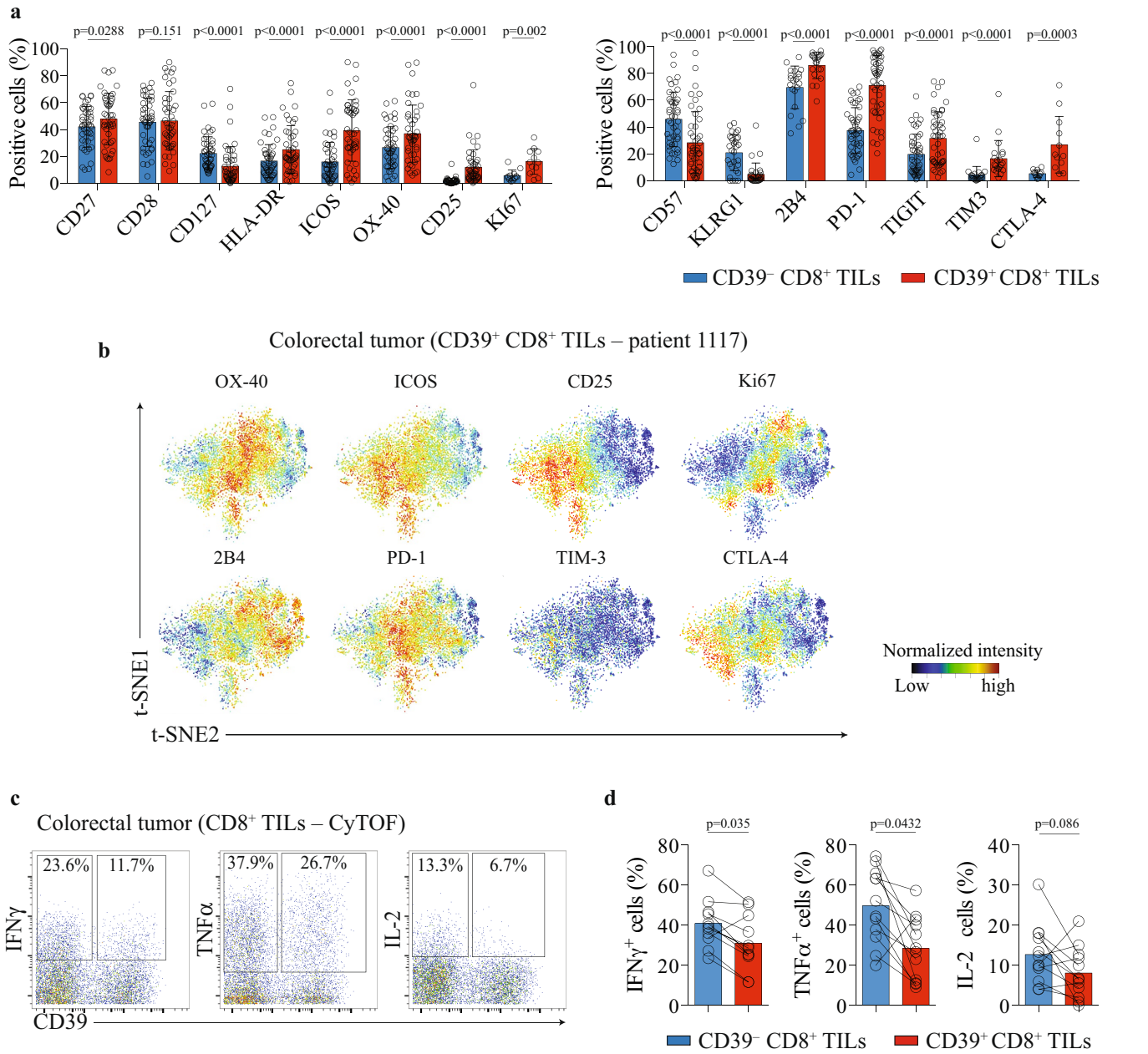
Extended Data Fig. 6 | Differential gene expression profiles of CD39⁻ and CD39⁺ CD8⁺ TILs. a, In order to better characterize CD39⁺ CD8⁺ TIL cells, we sorted and performed transcriptomic profiling on CD39⁻ and CD39⁺ CD8⁺ TILs. Using PCA on the complete transcriptomic data we observed a natural ordering of samples from naive to effector memory PBMCs, then CD39⁻ CD8⁺ TILs, and finally CD39⁺ CD8⁺ TILs along the PC1 axis (See Fig. 4). We then used GSEA to biologically interpret PC1. Among all pathways that were significantly upregulated, we found CD39⁺ CD8⁺ TILs were enriched in pathways related to cell proliferation and the adaptive immune response, which suggests that these cells were subjected to higher TCR signalling (See detailed list on Supplementary Table 5). **b**, To obtain a more comprehensive overview of the difference between CD39⁻ and CD39⁺ CD8⁺ TILs, we studied gene sets specific for exhaustion, a pathway characteristic of chronically stimulated T cells[28,29]. In line with the hypothesis that CD39 marks CD8⁺ TILs for chronic antigen stimulation, the gene set described for exhaustion (**c**) was significantly enriched in CD39⁺ CD8⁺ TILs in both colorectal and lung cancer (See Fig. 4).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



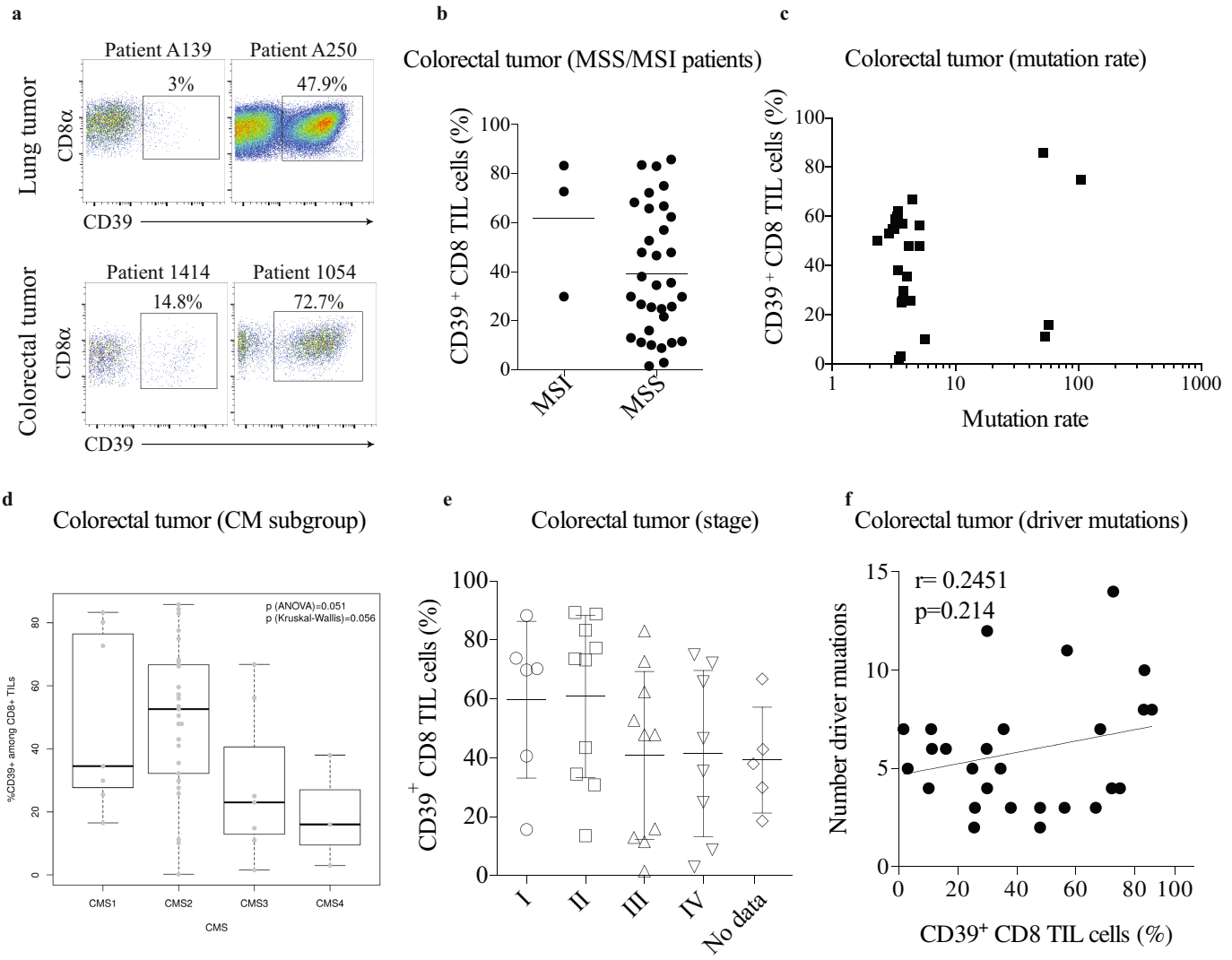
Extended Data Fig. 7 | Skewed TCR repertoire between CD39$^-$ and CD39$^+$ CD8$^+$ TILs. To further explore the specificity of CD39$^+$ CD8$^+$ TILs, we performed TCRα and TCRβ sequencing of CD39$^-$ and CD39$^+$ CD8$^+$ TILs. We assume that a less diverse TCRα or TCRβ profile in CD39$^+$ CD8$^+$ TILs would suggest tumour antigen-driven clonal expansion, as suggested[23]. **a**, The clonality index, incorporating the frequency of each unique TCRα or TCRβ clone in paired samples ($n = 8$ patients), indicated a lower TCRα and TCRβ diversity in CD39$^+$ CD8$^+$ TILs. Two-tailed paired *t*-test. Data are from two independent experiments. **b**, **c**, We also compared TCRα repertoires between these populations and found that the most highly represented clones were not shared between CD39$^-$ and CD39$^+$ CD8$^+$ TILs. Taken together, the less clonal and skewed TCRα profile of CD39$^+$ CD8$^+$ TILs supports the notion that these cells underwent tumour antigen-driven clonal expansion. Data are from two independent experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



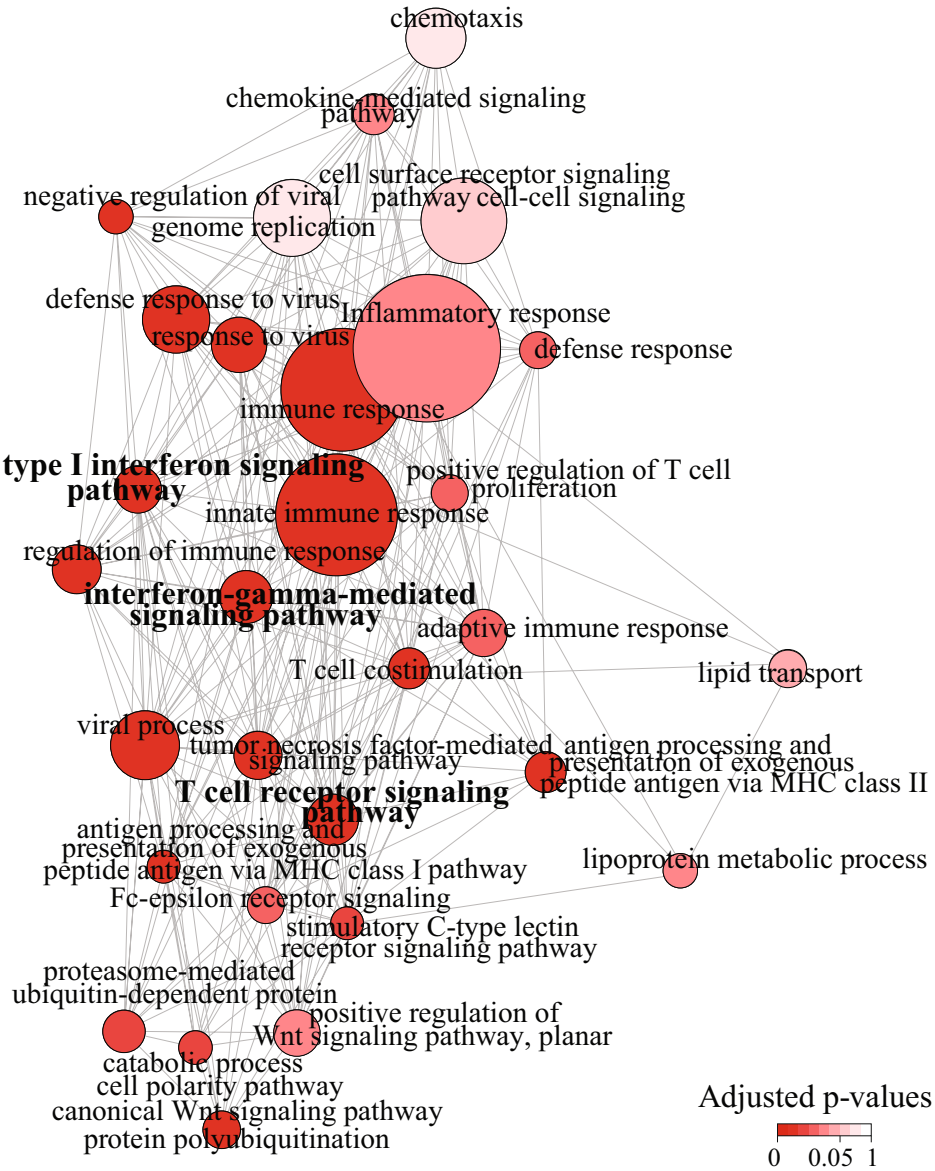
Extended Data Fig. 8 | Exhausted profiles of CD39$^+$ CD8$^+$ TILs. **a**, Frequencies of the expression of activation markers (left panel) and inhibitory markers (right panel) by CD39$^-$ (blue) and CD39$^+$ (red) CD8$^+$ TILs in lung tumours. Data are from at least ten independent mass cytometry experiments. Data are means ± s.d. Two-tailed paired *t*-test ($n = 12$ to $n = 30$ patients). **b**, *t*-SNE map of CD39$^+$ CD8$^+$ TILs cells isolated from a colorectal tumour. *t*-SNE was performed on data from one patient. Despite the phenotypic differences between CD39$^+$ and CD39$^-$ CD8$^+$ TILs, using t-SNE we observed that the CD39$^+$ CD8$^+$ TILs were heterogeneous and we could not describe any simple rules related to the hierarchical expression of various co-stimulatory receptors, inhibitory receptors or proliferation markers **c**, Mass cytometry dot plots representing expression of IFNγ, TNFα and IL-2 by CD8$^+$ TILs plotted against CD39 expression (representative data from one patient with a colorectal tumour). Data from two independent experiments. **d**, Frequency of the expression of IFNγ, TNFα and IL-2 by CD39$^-$ (blue) and CD39$^+$ (red) CD8$^+$ TILs from colorectal tumour. $n = 11$ patients, data from two independent experiments. Two-tailed paired *t*-test.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | Relationships between frequencies of CD39$^+$ CD8$^+$ TILs and clinical parameters of colorectal cancer patients. **a**, Mass cytometry dot plots representing expression of CD39 by CD8$^+$ TILs in lung and colorectal tumours. Representative data from two patients. Data are from at least ten independent mass cytometry experiments. **b**, CD39$^+$ cells as a percentage of CD8$^+$ TILs stratified by microsatellite-stable (MSS) ($n = 3$ patients) or microsatellite-instable (MSI) ($n = 33$ patients) status. Data are means from at least ten independent mass cytometry experiments. **c**, Mutation rate (in mutational events per megabase, plotted on a log scale) versus CD39$^+$ CD8$^+$ TIL frequencies in colorectal tumours ($n = 26$ patients). Data from at least ten independent mass cytometry experiments. **d**, Box plots representing CD39$^+$ frequencies among CD8$^+$ TILs stratified by the consensus molecular subtypes of each tumour. CMS1 ($n = 6$ patients), CMS2 ($n = 34$ patients), CMS3 ($n = 8$ patients), CMS4 ($n = 3$ patients). Box plots show the median, box edges represent the first and third quartiles, and the whiskers extend from minimum to maximum. Data are from at least ten independent mass cytometry experiments. **e**, Dot plots representing CD39$^+$ frequencies among CD8$^+$ TILs stratified by tumour stages. Stage I ($n = 6$ patients), stage II ($n = 10$ patients), stage III ($n = 10$ patients), stage IV ($n = 8$ patients). Data are mean ± s.d. from at least ten independent mass cytometry experiments. **f**, Number of driver mutations against CD39$^+$ CD8$^+$ TILs frequencies in colorectal tumours. Two-tailed *t*-test; Pearson's correlation. $n = 27$ patients. Data are from at least ten independent mass cytometry experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Gene set enrichment in tumours with high CD39$^+$ CD8$^+$ TIL count. We investigated transcriptomic profiles of whole tumours using bulk RNA sequencing in conjunction with the percentage CD39 expression in CD8$^+$ TILs as measured by mass cytometry. Among the 25% most varying genes, we identified ten gene modules by performing hierarchical clustering on the Pearson correlation matrix of the genes. Notably, a cluster whose expression correlated with the frequency of CD39$^+$ TILs was enriched in genes related to 'adaptive immune response', 'T cell receptor signalling pathway' and 'interferon-gamma mediated signalling pathway' (see also Supplementary Table 4). Pathways related to peptide presentation by MHC molecules were also overrepresented in this cluster, which contained genes such as class I MHC molecules, TAP1 and TAP2 molecules and proteasome-related genes. $n = 46$ patients. Data are from at least five independent experiments. Two-sided hypergeometric test.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0125-z

Microbial signals drive pre-leukaemic myeloproliferation in a *Tet2*-deficient host

Marlies Meisel[1,2,15], Reinhard Hinterleitner[1,2,15], Alain Pacis[3,4], Li Chen[1,2], Zachary M. Earley[1,2], Toufic Mayassi[1,2], Joseph F. Pierre[1,5], Jordan D. Ernest[1,2], Heather J. Galipeau[6], Nikolaus Thuille[7], Romain Bouziat[1,2], Manuel Buscarlet[8], Daina L. Ringus[1], Yitang Wang[9], Ye Li[9], Vu Dinh[1,2], Sangman M. Kim[1,2], Benjamin D. McDonald[1,9], Matthew A. Zurenski[1,2], Mark W. Musch[1], Glaucia C. Furtado[10], Sergio A. Lira[10], Gottfried Baier[7], Eugene B. Chang[1], A. Murat Eren[1,11], Christopher R. Weber[9], Lambert Busque[8,12], Lucy A. Godley[1,13], Elena F. Verdú[6], Luis B. Barreiro[3,14] & Bana Jabri[1,2,9]*

Somatic mutations in *tet methylcytosine dioxygenase 2* (*TET2*), which encodes an epigenetic modifier enzyme, drive the development of haematopoietic malignancies[1–7]. In both humans and mice, *TET2* deficiency leads to increased self-renewal of haematopoietic stem cells with a net developmental bias towards the myeloid lineage[1,4,8,9]. However, pre-leukaemic myeloproliferation (PMP) occurs in only a fraction of *Tet2*$^{-/-}$ mice[8,9] and humans with *TET2* mutations[1,3,5–7], suggesting that extrinsic non-cell-autonomous factors are required for disease onset. Here we show that bacterial translocation and increased interleukin-6 production, resulting from dysfunction of the small-intestinal barrier, are critical for the development of PMP in mice that lack *Tet2* expression in haematopoietic cells. Furthermore, in symptom-free *Tet2*$^{-/-}$ mice, PMP can be induced by disrupting intestinal barrier integrity, or in response to systemic bacterial stimuli such as the toll-like receptor 2 agonist. PMP was reversed by antibiotic treatment and failed to develop in germ-free *Tet2*$^{-/-}$ mice, which illustrates the importance of microbial signals in the development of this condition. Our findings demonstrate the requirement for microbial-dependent inflammation in the development of PMP and provide a mechanistic basis for the variation in PMP penetrance observed in *Tet2*$^{-/-}$ mice. This study will prompt new lines of investigation that may profoundly affect the prevention and management of haematopoietic malignancies.

TET2 deficiency leads to severe myeloproliferation, extramedullary haematopoiesis and splenomegaly that mimic pre-leukaemic myeloproliferative disorders (henceforth referred to as PMP)[2,7–9]. PMP is thought to set the stage for the development of overt leukaemia after the occurrence of additional cooperative oncogenic mutations[10,11]. To better understand the pathogenesis of PMP, we set out to determine the mechanisms that underlie variations in PMP penetrance, which are observed in around 50–75% of aged *Tet2*$^{-/-}$ mice that are over 20 weeks old[8,9] (Extended Data Fig. 1a). We first developed a peripheral-blood biomarker that could serve as a robust predictor of disease. We found that the frequency of CD11b^{+}Gr1^{+} myeloid cells in the peripheral blood positively correlates with myeloid proliferations[12] (as defined by the expansion of myeloid cells in the spleen and peripheral blood) and splenic extramedullary haematopoiesis markers[8,9,13,14] (an increase of lineage^{-}Sca1^{+}c-Kit^{+} (LSK) cells and splenomegaly) (Extended Data Fig. 1b–e). Confirming the prognostic robustness of this newly defined peripheral-blood biomarker of PMP, *Tet2*$^{-/-}$ mice that were over 20 weeks old and had >16% of CD11b^{+}Gr1^{+} myeloid cells in their peripheral blood displayed a PMP phenotype (Extended Data Fig. 1f–h) as

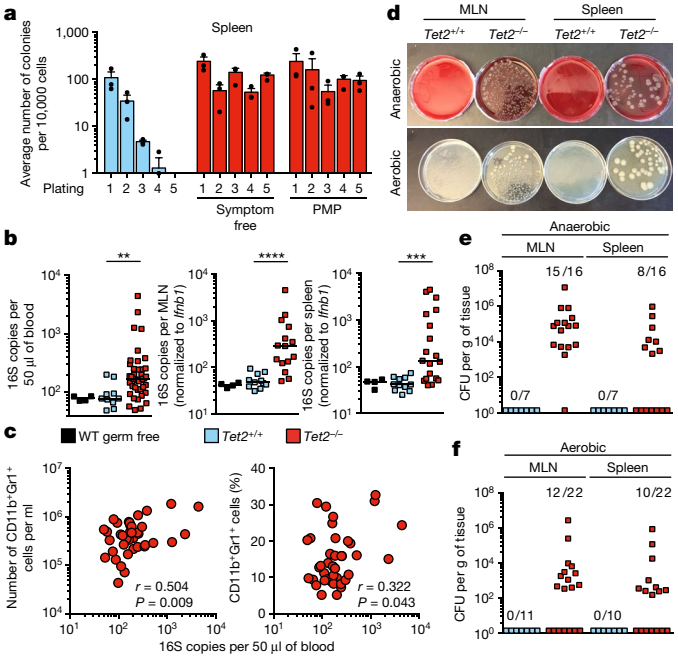


Fig. 1 | *Tet2* deficiency leads to systemic bacterial dissemination. **a**, In vitro HSC self-renewal colony-forming assay of haematopoietic progenitors ($n = 3$ mice). Mean ± s.e.m. **b**, Bacterial 16S gene copies in peripheral blood (left; $n = 4$, 10 and 37 for germ-free wild-type (WT), *Tet2*$^{+/+}$ and *Tet2*$^{-/-}$ mice, respectively), MLN (middle; $n = 4$, 11 and 15 for germ-free wild-type, *Tet2*$^{+/+}$ and *Tet2*$^{-/-}$ mice, respectively) and spleen (right; $n = 4$, 11 and 19 for germ-free wild-type, *Tet2*$^{+/+}$ and *Tet2*$^{-/-}$ mice, respectively). Germ-free wild-type mice served as negative control. Centre at median, two-tailed Mann–Whitney *U*-test. Bacterial 16S rRNA gene qPCRs from about 30 mg MLN and about 30 mg spleen were normalized to the host murine *Ifnb1* gene. **c**, Correlation between 16S gene copies in the peripheral blood and numbers (left) or frequency (right) of peripheral-blood CD11b^{+}Gr1^{+}myeloid cells ($n = 40$ mice). Pearson correlation test. **d**, Representative image of aerobic and anaerobic cultures. **e**, **f**, Quantification of bacteria colony-forming units (CFUs) of MLN and spleen suspensions grown under anaerobic (**e**) or aerobic (**f**) conditions. In **b**, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. Data are representative of at least three independent experiments.

[1]Department of Medicine, University of Chicago, Chicago, IL, USA. [2]Committee on Immunology, University of Chicago, Chicago, IL, USA. [3]Department of Genetics, CHU Sainte-Justine Research Center, Montreal, Quebec, Canada. [4]Department of Biochemistry, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada. [5]Department of Pediatrics, University of Tennessee Health Science Center, Memphis, TN, USA. [6]Farncombe Family Digestive Health Research Institute, McMaster University, Hamilton, Ontario, Canada. [7]Translational Cell Genetics, Department for Pharmacology and Genetics, Medical University of Innsbruck, Innsbruck, Austria. [8]Research Centre, Hôpital Maisonneuve-Rosemont, Montreal, Quebec, Canada. [9]Department of Pathology and Pediatrics, University of Chicago, Chicago, IL, USA. [10]Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [11]Marine Biological Laboratory, Woods Hole, MA, USA. [12]Hematology Division, Hôpital Maisonneuve-Rosemont, Université de Montréal, Montreal, Quebec, Canada. [13]Section of Hematology/Oncology, University of Chicago, Chicago, IL, USA. [14]Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada. [15]These authors contributed equally: Marlies Meisel, Reinhard Hinterleitner. *e-mail: bjabri@bsd.uchicago.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

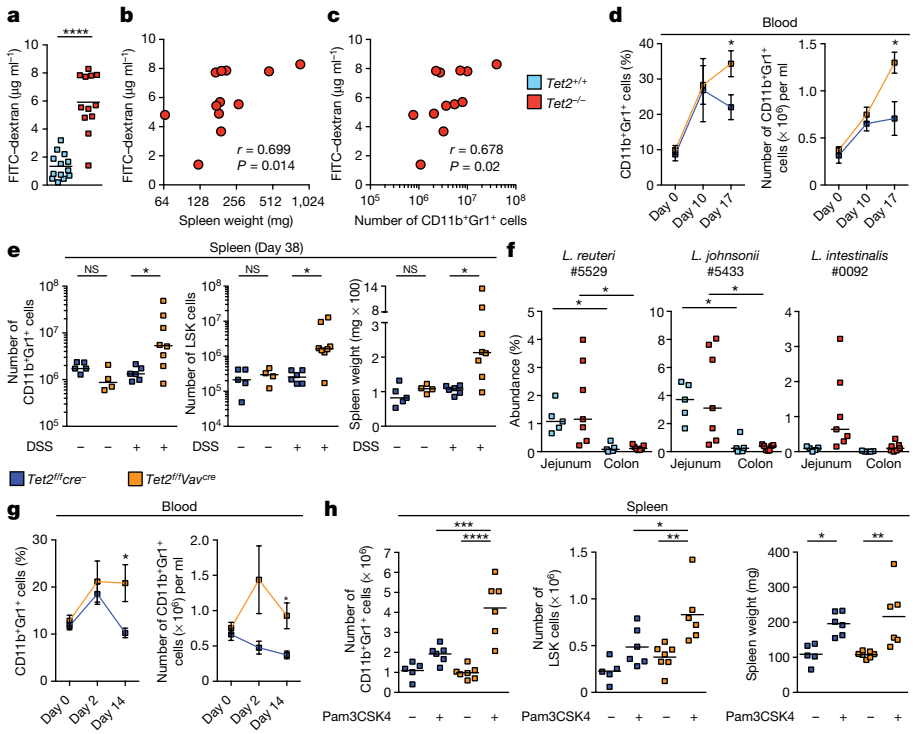


Fig. 2 | Disruption of intestinal barrier integrity and systemic microbial signals are sufficient to drive PMP in the absence of *Tet2*. **a**, Intestinal permeability measurement by blood plasma FITC–dextran concentrations ($n = 13$ $Tet2^{+/+}$ mice or 12 $Tet2^{-/-}$ mice). Centre is mean, two-tailed unpaired *t*-test. **b**, **c**, Blood plasma FITC–dextran concentration correlates with spleen weight (**b**) and numbers of splenic $CD11b^+Gr1^+$ myeloid cells (**c**) ($n = 12$ mice). Pearson correlation test. **d**, **e**, Symptom-free $Tet2^{f/f}Vav^{cre}$ mice that are over 20 weeks old, and littermate controls, treated with DSS (+) or received water (−). **d**, Frequency (left) and numbers (right) of $CD11b^+Gr1^+$ myeloid cells before, during and after DSS treatment ($n = 5$ $Tet2^{f/f}cre^-$ ($Tet2^{f/f}$ mice without the Cre construct) mice or 8 $Tet2^{f/f}Vav^{cre}$ mice). Mean ± s.e.m., two-tailed unpaired *t*-test. **e**, Numbers of $CD11b^+Gr1^+$ myeloid cells (left), LSK cells (middle) and spleen weight (right) at end point analysis ($n = 5$ ($Tet2^{f/f}cre^-$, no DSS), 4 ($Tet2^{f/f}Vav^{cre}$, no DSS), 6 ($Tet2^{f/f}cre^-$, DSS) or 8 ($Tet2^{f/f}Vav^{cre}$, DSS) mice). **f**, Oligotypes identified in MLNs and spleens of $Tet2^{-/-}$ mice are enriched in the jejunum. Centre is median ($n = 5$ ($Tet2^{+/+}$) or 7 ($Tet2^{-/-}$) mice for jejunum and colon). **g**, **h**, Symptom-free $Tet2^{f/f}Vav^{cre}$ that are over 20 weeks old, and littermate controls, treated with the TLR1/2 agonist Pam3CSK4 (+) or received PBS (−). **g**, Percentage (left) and numbers (right) of $CD11b^+Gr1^+$ myeloid cells prior (day 0), during treatment (day 2) and at end-point analysis (day 14) ($n = 6$ mice). Mean ± s.e.m., two-tailed paired *t*-test, Sidak's post hoc test. **h**, Numbers of $CD11b^+Gr1^+$ myeloid cells (left), LSK cells (middle) and spleen weight (right) ($n = 5$ ($Tet2^{f/f}cre^-$, no Pam3CSK4), 6 ($Tet2^{f/f}cre^-$, with Pam3CSK4), 7 ($Tet2^{f/f}Vav^{cre}$, no Pam3CSK4) or 6 ($Tet2^{f/f}Vav^{cre}$, with Pam3CSK4) mice). Centre is mean, one-way ANOVA, Sidak's post hoc test. In **e**, **f**, centre is median, Kruskal–Wallis, Dunn's post hoc test. Data are representative of at least two independent experiments. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

compared to age-matched $Tet2^{-/-}$ mice with ≤16% peripheral-blood $CD11b^+Gr1^+$ cells (defined as 'symptom free'), or littermate controls.

We next determined whether differences in PMP penetrance in $Tet2^{-/-}$ mice (Extended Data Fig. 1a–h) could be linked to a change in the haematopoietic stem cell (HSC) self-renewing potential in vitro[8,9]. Irrespective of health status, bone marrow and splenic haematopoietic progenitors from symptom-free $Tet2^{-/-}$ mice and $Tet2^{-/-}$ mice with PMP displayed similar re-plating efficiencies, which were superior to those of littermate controls (Fig. 1a and Extended Data Fig. 1i). This suggests that extrinsic non-cell-autonomous factors are required for the development of PMP in $Tet2^{-/-}$ mice.

This finding, together with previous studies that report that the systemic dissemination of bacteria induces extramedullary haematopoiesis and emergency myelopoiesis[15], prompted us to investigate whether bacterial triggers could drive myeloproliferation in the absence of *Tet2*. Quantification of the bacterial 16S rRNA gene revealed the presence of 16S gene copies in the peripheral blood, the mesenteric lymph nodes (MLNs), and the spleen of $Tet2^{-/-}$ mice (Fig. 1b) that correlated with myelocytosis (Fig. 1c and Extended Data Fig. 1j, k). Furthermore, cultivation of MLN and spleen suspensions under aerobic and anaerobic culture conditions confirmed that in about 50% of $Tet2^{-/-}$ mice, live bacteria—in particular *Lactobacillus*—were present. *Lactobacillus* has previously been reported to be predominantly located in the small intestine under steady-state conditions[16] (Fig. 1d–f, and Extended Data Fig. 1l). No bacterial dissemination was observed in littermate controls (Fig. 1d–f).

The existence of spontaneous bacterial translocation in $Tet2^{-/-}$ mice was unexpected, given the absence of intestinal anomalies (Extended Data Fig. 2a) and epithelial cell death (Extended Data Fig. 2b, c) in $Tet2^{-/-}$ mice. However, increased intestinal permeability can occur in mice and humans with normal intestinal architecture[17]. Consistent with the observed microbial translocation, $Tet2^{-/-}$ mice displayed a significant increase in intestinal permeability as assessed by in vivo FITC–dextran intestinal permeability assay (Fig. 2a). Increased intestinal permeability positively correlated with PMP severity in $Tet2^{-/-}$ mice (Fig. 2b, c and Extended Data Fig. 3a), suggesting that disruption of intestinal barrier integrity may be sufficient to initiate PMP in symptom-free $Tet2^{-/-}$ mice. Accordingly, administration of dextran sodium sulfate (DSS)—a compound known to alter intestinal barrier function[18]—caused excessive myeloproliferation and extramedullary haematopoiesis that persisted more than one month after cessation of treatment in symptom-free $Tet2^{f/f}Vav^{cre}$ mice, but not in littermate controls (Fig. 2d, e and Extended Data Fig. 3b–d). $Tet2^{f/f}Vav^{cre}$ mice (*Tet2* deleted in haematopoietic cells including progenitors) were used because somatic *TET2* mutations occur in the haematopoietic compartment in humans[1–7]. These results suggest that maintaining intestinal barrier function is important for preventing PMP in the context of *Tet2* deficiency in haematopoietic cells.

Whole transcriptome sequencing analysis of the intestine of $Tet2^{-/-}$ and littermate control mice revealed major transcriptional alterations in the jejunum, but minor changes were also detected in the colon (Extended Data Fig. 4a). This finding is consistent with the

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

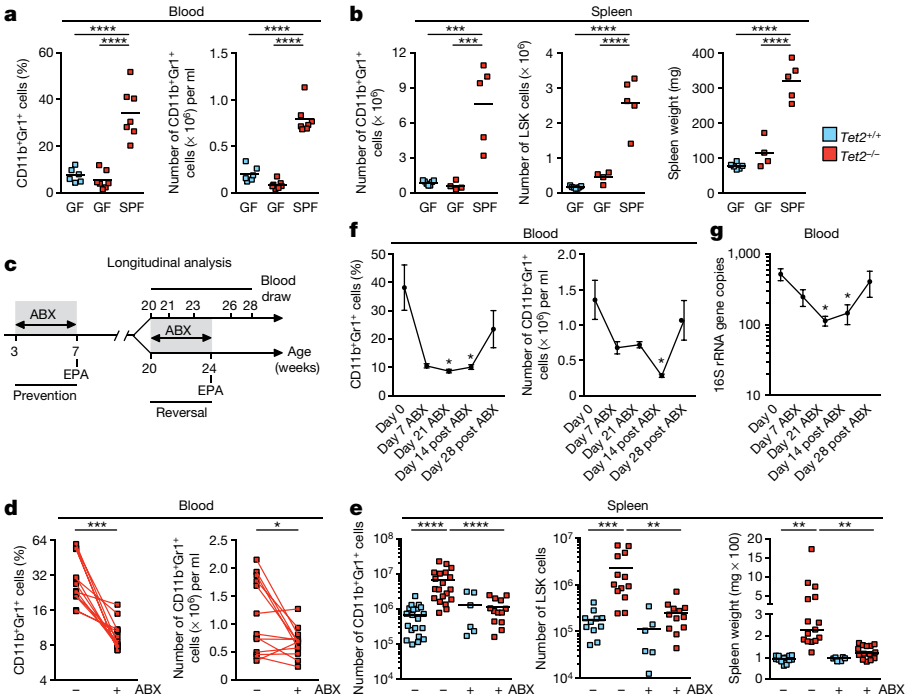


Fig. 3 | Microbial signals are required for PMP in *Tet2*$^{-/-}$ mice. **a**, **b**, Germ-free (GF), SPF-housed (SPF) *Tet2*$^{-/-}$ mice that are over 40 weeks old, and littermate controls, analysed for percentage (**a**, left) and numbers (**a**, right) of CD11b^{+}Gr1^{+} myeloid cells ($n = 6$ (GF *Tet2*$^{+/+}$), 7 (GF *Tet2*$^{-/-}$) or 7 (SPF *Tet2*$^{-/-}$) mice), and numbers of CD11b^{+}Gr1^{+} myeloid cells (**b**, left), LSK cells (**b**, middle), and spleen weight (**b**, right), ($n = 7$ (GF *Tet2*$^{+/+}$), 4 (GF *Tet2*$^{-/-}$) or 5 (SPF *Tet2*$^{-/-}$) mice). **c**, Schematic for antibiotics (ABX) treatment regimen. Longitudinal analysis described in **f**, **g** and Extended Data Fig. 8d; prevention described in Extended Data Fig. 8c; and reversal described in **d**, **e** and Extended Data Fig. 9a–j. EPA, end point analysis. **d**, **e**, *Tet2*$^{-/-}$ mice with PMP that are over 20 weeks old, and littermate controls, analysed for percentage (**d**, left) and numbers (**d**, right) of CD11b^{+}Gr1^{+} myeloid cells ($n = 13$ mice). Lines connect values obtained from the same mouse sampled before (−) and after (+) ABX treatment. Two-tailed paired *t*-test. **e**, Numbers of CD11b^{+}Gr1^{+} myeloid cells (left; $n = 20$ (*Tet2*$^{+/+}$, no ABX treatment), 20 (*Tet2*$^{-/-}$, no ABX treatment), 6 (*Tet2*$^{+/+}$, with ABX treatment) and 13 (*Tet2*$^{-/-}$, with ABX treatment) mice), LSK cells (middle; $n = 11$ (*Tet2*$^{+/+}$, no ABX treatment), 13 (*Tet2*$^{-/-}$, no ABX treatment), 6 (*Tet2*$^{+/+}$, with ABX treatment) and 12 (*Tet2*$^{-/-}$, with ABX treatment) mice) and spleen weight (right; $n = 14$ (*Tet2*$^{+/+}$, no ABX treatment), 15 (*Tet2*$^{-/-}$, no ABX treatment), 6 (*Tet2*$^{+/+}$, with ABX treatment) and 14 (*Tet2*$^{-/-}$, with ABX treatment) mice). **f**, **g**, *Tet2*$^{-/-}$ mice analysed for frequency (**f**, left) and number (**f**, right) of CD11b^{+}Gr1^{+} myeloid cells and (**g**) 16S gene copies before, during and after ABX treatment ($n = 7$ mice). Mean ± s.e.m., repeated measures one-way ANOVA, Sidak's post hoc test. **a**, **b**, **e**, Centre is mean, one-way ANOVA, Sidak's post hoc test. Data are representative of at least two independent experiments; $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

observation that *Lactobacillus* spp. residing in the small intestine[16] were predominantly found in peripheral organs of *Tet2*$^{-/-}$ mice. Consistent with *Tet2*-deficient mice displaying increased intestinal permeability (Fig. 2a), the expression of genes known to maintain intestinal immune homeostasis and barrier function was altered in the jejunum (Extended Data Fig. 4b). In particular, tight junction zonula occludens-1 (ZO-1), which is critical in regulating the paracellular leakage pathway[19], was reduced in the jejunum but not in the colon of *Tet2*-deficient mice (Extended Data Fig. 4c-d). In contrast to previous observations that suggested that selective deletion of *Tet1* in epithelial cells leads to defects in intestinal permeability[20], *Tet2*$^{f/f}$*Vav*^{cre} mice—but not *Tet2*$^{f/f}$ *Villin*^{cre} mice (*Tet2* deleted in epithelial cells)—displayed defects in gut barrier function (Extended Data Fig. 5a–c). Furthermore, *Tet2*$^{f/f}$ *LysM*^{cre} (*LysM* is also known as *Lyz2*) mice (*Tet2* deleted in mature myeloid cells) failed to show an increase in intestinal permeability (Extended Data Fig. 5a–c). Consistent with these findings, and as previously shown[8,21], PMP developed only in *Tet2*$^{f/f}$*Vav*^{cre} mice (Extended Data Fig. 5d–f). These observations suggest that *Tet2* deficiency in the haematopoietic compartment can promote small-intestinal barrier dysfunction that causes bacterial translocation.

To further assess whether the translocated bacteria were primarily of jejunal origin, we matched oligotypes[22] of identified bacterial strains in the spleen and MLN of *Tet2*$^{-/-}$ mice with bacteria residing in the different intestinal compartments (Extended Data Fig. 1l and Supplementary Table 4). Consistent with the proposed hypothesis, *Lactobacillus reuteri* #5529 (number refers to oligotype), *Lactobacillus johnsonii* #5433 and *Lactobacillus intestinalis* #0092 were substantially enriched in the jejunum independent of the genotype, when compared to the colon (Fig. 2f).

We first assessed whether TLR2 agonists—cell wall components of several *Lactobacillus* strains[23]—are relevant bacterial signals in *Tet2*$^{-/-}$ mice (Extended Data Fig. 6a), and then showed that the TLR2 agonist Pam3CSK4 was sufficient to induce extensive myeloproliferation in symptom-free *Tet2*$^{f/f}$*Vav*^{cre} mice but not in littermate controls (Fig. 2g–h and Extended Data Fig. 6b–d), without affecting intestinal barrier integrity (Extended Data Fig. 6e). Of note, *L. reuteri* #5529 and *L. johnsonii* #5433 are also found in the small intestine of littermate control mice (Fig. 2f) and 16S microbiome profiling failed to detect significant differences in microbial structures between genotypes (Extended Data Fig. 7a–d). Furthermore, variations in disease penetrance were observed among cohoused *Tet2*$^{f/f}$*Vav*^{cre} mice (Extended Data Fig. 7e). These observations indicate that changes in microbial structures do not drive PMP and that systemic microbial signals, such as TLR2 agonists, may be sufficient to promote PMP independently of intestinal barrier dysfunction in *Tet2*-deficient mice. This is consistent with previous epidemiological studies that report an association between clonal expansion and myeloid malignancies with antecedent chronic inflammatory conditions of infectious origin[24].

To demonstrate that microbial signals are required for the development of PMP, we raised *Tet2*$^{-/-}$ mice under germ-free conditions. Forty-week-old *Tet2*$^{-/-}$ germ-free mice failed to show signs of PMP in the peripheral blood and the spleen, in contrast to age-matched specific-pathogen-free (SPF)-raised *Tet2*$^{-/-}$ mice (Fig. 3a, b and Extended Data Fig. 8a, b). Furthermore, treatment with antibiotics (Fig. 3c) both

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

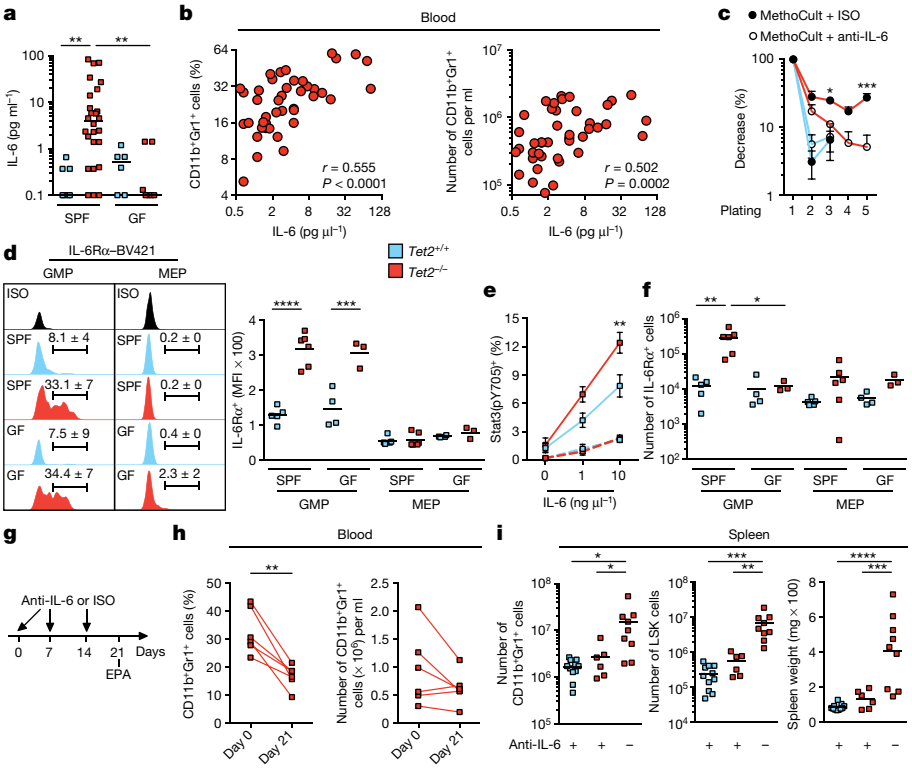


Fig. 4 | Bacteria-induced IL-6 is required for PMP in *Tet2*$^{-/-}$ mice. **a**, IL-6 cytokine levels in blood plasma ($n = 7$ (*Tet2*$^{+/+}$, SPF), 26 (*Tet2*$^{-/-}$, SPF), 6 (*Tet2*$^{+/+}$, GF) or 7 (*Tet2*$^{-/-}$, GF) mice). Centre is median, Kruskal–Wallis, Dunn's post hoc test. **b**, IL-6 cytokine levels in blood plasma correlate with the frequency (left) and numbers (right) of CD11b^{+}Gr1^{+} myeloid cells ($n = 49$ mice). Pearson correlation test. **c**, In vitro HSC self-renewal colony-forming assay of haematopoietic progenitor of the spleen from *Tet2*$^{-/-}$ mice (red lines) and littermate controls (blue lines) in the presence of anti-IL-6 antibody (anti-IL-6) or isotype control (ISO) after the first replating ($n = 3$ mice). **d**, Representative histograms, percentages and quantification of mean fluorescence intensity (MFI) of IL-6Rα^{+}c-Kit^{+}Sca-1^{-} (lineage^{-}c-Kit^{+} (LK) gated) CD34^{+}FcγRIII/II^{+} (FcγRIII and FcγRII are also known as CD16 and CD32, respectively) GMPs and IL-6Rα^{+} LK gated CD34^{-}FcγRIII/II^{-} megakaryocyte–erythroid progenitors (MEP) from the spleen ($n = 5$ (*Tet2*$^{+/+}$, SPF), 6 (*Tet2*$^{-/-}$, SPF), 4 (*Tet2*$^{+/+}$, GF) or 3 (*Tet2*$^{-/-}$, GF) mice). **e**, Stat3 phosphorylation (pY705) after 30-min stimulation with IL-6 in splenic c-Kit^{+}Sca-1^{-} (LK gated) CD34^{+} (solid lines) and CD34^{-} (dotted lines) myeloid progenitors ($n = 5$ (blue line), 6 (red line) and 4 mice (dotted lines)). **f**, Numbers of IL-6Rα^{+} GMP and megakaryocyte–erythroid progenitors from the spleen ($n = 5$ (*Tet2*$^{+/+}$, SPF), 6 (*Tet2*$^{-/-}$, SPF), 4 (*Tet2*$^{+/+}$, GF) or 3 (*Tet2*$^{-/-}$, GF) mice). **g**, Schematic for anti-IL-6 treatment of *Tet2*$^{-/-}$ mice with PMP that are over 20 weeks old, and littermate controls. See data in **h**, **i** and Extended Data Fig. 10o, p. **h**, Percentage (left) and numbers (right) of CD11b^{+}Gr1^{+} myeloid cells at day 0 and day 21 ($n = 6$ mice). Lines connect values obtained from the same mouse sampled before and after anti-IL-6 treatment. Two-tailed paired *t*-test. **i**, Numbers of CD11b^{+}Gr1^{+} myeloid cells (left), LSK cells (middle) and spleen weight (right) ($n = 11$ (*Tet2*$^{+/+}$, with anti-IL-6), 6 (*Tet2*$^{-/-}$, with anti-IL-6) or 9 (*Tet2*$^{-/-}$, without anti-IL-6) mice). **c**, **e**, Mean ± s.e.m., two-way ANOVA, Sidak's post hoc test. **d**, **f**, **i**, Centre is mean, one-way ANOVA, Sidak's post hoc test. Data are representative of at least two independent experiments. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

prevented (Extended Data Fig. 8c) and reversed PMP (Fig. 3d, e and Extended Data Fig. 9a–j) in *Tet2*$^{-/-}$ mice. The longitudinal study that was performed in the presence of antibiotics and after antibiotic withdrawal (Fig. 3c), using the peripheral-blood PMP marker (Fig. 3f, g and Extended Data Fig. 8d), indicated that the potential for PMP remains and is directly linked to the presence of bacterial signals.

Interleukin 6 (IL-6) is a critical activator of myelopoiesis in response to systemic bacterial dissemination[15] and can be upregulated in chronic myeloproliferative disease in humans[25]. Accordingly, IL-6 was significantly increased in both the plasma and spleen of *Tet2*$^{-/-}$ mice with PMP, as compared to littermate controls (Fig. 4a and Extended Data Fig. 10a). The increase in IL-6 was microbiota-dependent and induced upon DSS and TLR2-agonist treatment in symptom-free *Tet2*-deficient mice that subsequently developed PMP (Fig. 4a and Extended Data Fig. 10a–d). Furthermore, peripheral-blood myeloid cell expansion correlated with IL-6 levels in plasma (Fig. 4b and Extended Data Fig. 10e). We next investigated the role of IL-6 in PMP in the context of *Tet2* deficiency in vitro and in vivo. Of note, IL-6 is used in in vitro HSC self-renewing assays (Fig. 1a and Extended Data Fig. 1i[26]). The neutralization of IL-6 significantly reduced the increased self-renewing capacity of *Tet2*$^{-/-}$ haematopoietic progenitors in vitro, which indicates a critical role for IL-6 in PMP, but IL-6 neutralization did not affect the limited self-renewing capacity of *Tet2*-sufficient haematopoietic progenitors (Fig. 4c and Extended Data Fig. 10f, g). Consistent with this finding, IL-6Rα surface expression was significantly increased on *Tet2*$^{-/-}$ granulocyte–macrophage progenitors (GMPs) as compared to *Tet2*$^{+/+}$ GMPs (Fig. 4d and Extended Data Fig. 10h) and associated with an increased sensitivity of *Tet2*$^{-/-}$ GMPs to IL-6 as assessed by Stat3 phosphorylation (Fig. 4e and Extended Data Fig. 10i). Although an increase in IL-6Rα expression was observed in germ-free *Tet2*$^{-/-}$ GMPs, the cellular expansion of IL-6Rα^{+} GMPs occurred only in SPF-housed *Tet2*$^{-/-}$ mice (Fig. 4f). This suggests that the increase in IL-6Rα expression is intrinsic to the cell and independent of microbial signals, whereas the expansion of IL-6Rα^{+} GMPs requires the microbiota and IL-6 (Extended Data Fig. 8b). Importantly, megakaryocyte–erythroid progenitors did not display increased IL-6Rα expression and signalling (Fig. 4d, e) and consequently failed to expand in *Tet2*$^{-/-}$ mice under SPF conditions (Fig. 4f). These data suggest that the development of PMP requires an exogenous microbial signal and an upregulation of IL-6Rα on GMPs that is driven in a cell-autonomous manner by *Tet2* deficiency. Of note, increased IL-6 production in the plasma and spleen (Fig. 4a and Extended Data Fig. 10a) of *Tet2*$^{-/-}$ mice correlated with an increase in the number of myeloid cells capable of producing IL-6 (Extended Data Fig. 10j), rather than an increase in IL-6 production by mature myeloid cells (Extended Data Fig. 10k, l). No increase in IL-6Rα expression in mature splenic *Tet2*$^{-/-}$ CD11b^{+}Gr1^{+} myeloid

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

cells was observed (Extended Data Fig. 10m), which suggests that IL-6-mediated PMP in *Tet2*-deficient mice is determined at the myeloid progenitor stage. Consistent with this observation, $Tet2^{f/f}Vav^{cre}$ but not $Tet2^{f/f}LysM^{cre}$ mice developed PMP (Extended Data Figs. 5d–f, 10n). Finally, in vivo IL-6 neutralization experiments established that IL-6 was required for PMP to develop in $Tet2^{-/-}$ mice (Fig. 4g–i and Extended Data Fig. 10o) independently of barrier function restoration (Extended Data Fig. 10p, q). These findings provide experimental support to studies in humans that suggest a role for IL-6 in the pathogenesis of haematological malignancies[27].

Our study demonstrates a critical role for microbial-mediated inflammatory (IL-6) signals in the development of PMP in the context of *Tet2* deficiency. More specifically, these data suggest an IL-6 dependent cycle that is engaged upon bacterial translocation and leads to PMP in $Tet2^{-/-}$ mice through the expansion of GMPs that express high levels of IL-6Rα in the absence of *Tet2* (Extended Data Fig. 10r for a schematic of this model). The mechanisms through which *Tet2* deficiency in haematopoietic cells leads to a microbiota-dependent impairment of gut barrier function remain to be addressed.

This work raises the question of whether microbial-dependent inflammatory mediators, such as IL-6, are critical contributors to myeloid malignancies in humans by promoting PMP in patients with somatic *TET2* mutations. Furthermore, our findings—in combination with a report indicating that the restoration of *Tet2* expression after PMP has developed prevents secondary evolution towards leukaemia in $Tet2^{-/-}$ mice[28]—suggest that blocking inflammatory bacterial signals in patients with *TET2* deficiency and PMP may reduce the risk of progression to haematopoietic malignancies. Whether bacterial infections and events promoting the disruption of intestinal barrier function can create a permissive environment that evokes the acquisition of cooperative oncogenic mutations that lead to the development of leukaemia[10] remains to be determined. Irrespectively, our study will motivate a new line of investigations and the design of novel therapeutic strategies that target IL-6 in patients with PMP linked to somatic *TET2* mutations to revert myeloproliferation and prevent the development of myeloid malignancies.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0125-z.

Received: 1 June 2017; Accepted: 17 April 2018;
Published online: 16 May 2018

1. Delhommeau, F. et al. Mutation in *TET2* in myeloid cancers. *N. Engl. J. Med.* **360**, 2289–2301 (2009).
2. Kosmider, O. et al. *TET2* gene mutation is a frequent and adverse event in chronic myelomonocytic leukemia. *Haematologica* **94**, 1676–1681 (2009).
3. Busque, L. et al. Recurrent somatic *TET2* mutations in normal elderly individuals with clonal hematopoiesis. *Nat. Genet.* **44**, 1179–1181 (2012).
4. Abdel-Wahab, O. & Levine, R. L. Mutations in epigenetic modifiers in the pathogenesis and therapy of acute myeloid leukemia. *Blood* **121**, 3563–3572 (2013).
5. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
6. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
7. Jan, M., Ebert, B. L. & Jaiswal, S. Clonal hematopoiesis. *Semin. Hematol.* **54**, 43–50 (2017).
8. Moran-Crusio, K. et al. *Tet2* loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11–24 (2011).
9. Quivoron, C. et al. TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer Cell* **20**, 25–38 (2011).
10. Bowman, R. L., Busque, L. & Levine, R. L. Clonal hematopoiesis and evolution to hematopoietic malignancies. *Cell Stem Cell* **22**, 157–170 (2018).
11. Kunimoto, H. et al. Cooperative epigenetic remodeling by TET2 loss and NRAS mutation drives myeloid transformation and MEK inhibitor sensitivity. *Cancer Cell* **33**, 44–59.e48 (2018).
12. Kogan, S. C. et al. Bethesda proposals for classification of nonlymphoid hematopoietic neoplasms in mice. *Blood* **100**, 238–245 (2002).
13. Ko, M. et al. Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. *Proc. Natl Acad. Sci. USA* **108**, 14566–14571 (2011).
14. Li, Z. et al. Deletion of *Tet2* in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies. *Blood* **118**, 4509–4518 (2011).
15. Boettcher, S. & Manz, M. G. Regulation of inflammation- and infection-driven hematopoiesis. *Trends Immunol.* **38**, 345–357 (2017).
16. Sartor, R. B. Microbial influences in inflammatory bowel diseases. *Gastroenterology* **134**, 577–594 (2008).
17. Marchiando, A. M., Graham, W. V. & Turner, J. R. Epithelial barriers in homeostasis and disease. *Annu. Rev. Pathol.* **5**, 119–144 (2010).
18. Chassaing, B., Aitken, J. D., Malleshappa, M. & Vijay-Kumar, M. Dextran sulfate sodium (DSS)-induced colitis in mice. *Curr. Protoc. Immunol.* **104**, 15.25.1–15.25.14 (2014).
19. Shen, L., Weber, C. R., Raleigh, D. R., Yu, D. & Turner, J. R. Tight junction pore and leak pathways: a dynamic duo. *Annu. Rev. Physiol.* **73**, 283–309 (2011).
20. Chapman, C. G. et al. TET-catalyzed 5-hydroxymethylcytosine regulates gene expression in differentiating colonocytes and colon cancer. *Sci. Rep.* **5**, 17568 (2015).
21. Zhao, Z. et al. The catalytic activity of TET2 is essential for its myeloid malignancy-suppressive function in hematopoietic stem/progenitor cells. *Leukemia* **30**, 1784–1788 (2016).
22. Eren, A. M. et al. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* **4**, 1111–1119 (2013).
23. Wells, J. M. Immunomodulatory mechanisms of lactobacilli. *Microb. Cell Fact.* **10**, S17 (2011).
24. Kristinsson, S. Y. et al. Chronic immune stimulation might act as a trigger for the development of acute myeloid leukemia or myelodysplastic syndromes. *J. Clin. Oncol.* **29**, 2897–2903 (2011).
25. Panteli, K. E. et al. Serum interleukin (IL)-1, IL-2, sIL-2Ra, IL-6 and thrombopoietin levels in patients with chronic myeloproliferative diseases. *Br. J. Haematol.* **130**, 709–715 (2005).
26. Miller, C. L. & Lai, B. Human and mouse hematopoietic colony-forming cell assays. *Methods Mol. Biol.* **290**, 71–89 (2005).
27. Johnson, D. E., O'Keefe, R. A. & Grandis, J. R. Targeting the IL-6/JAK/STAT3 signalling axis in cancer. *Nat. Rev. Clin. Oncol.* **15**, 234–248 (2018).
28. Cimmino, L. et al. Restoration of TET2 function blocks aberrant self-renewal and leukemia progression. *Cell* **170**, 1079–1095.e1020 (2017).

Acknowledgements We thank the University of Chicago DNA Sequencing Core facility for assistance with sequencing; the Calcul Québec and Compute Canada for providing access to the supercomputer Briaree from the University of Montreal; the Histology Core facility at the University of Chicago Human Tissue Resource Center for assistance with histology; S. Hwang (University of Chicago) for providing LysM-Cre mice; and V. Abadie (Université de Montréal) for discussions and critical reading of the manuscript. This work was supported by grants from the Cancer Center Support Grant P30CA014599 to B.J.; Digestive Diseases Research Core Center P30DK42086 at the University of Chicago to R.H., E.B.C., C.R.W. and B.J.; P01DK072201, R01DK110352 and 5R01CA161373 to S.A.L. and G.C.F.; F32 DK105728-01A1 to J.F.P.; CCFA Research Fellowship Award (ID: 480735) to M.M.; FWF Austrian Science Fund (P30324-B21) and Christian Doppler Society (I-CARE) to G.B.; and by a CIHR grant MOP (#20003029), a Canada Research Chair to E.F.V. and a Canada Research Chair to L.B.B. A.P. was supported by a fellowship from FRQS.

Reviewer information *Nature* thanks S. Grivennikov, K. Shannon, Y. Xiong and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.M., R.H. and B.J. conceived and designed the research, guided the interpretation of the results and wrote the manuscript. M.M. and R.H. performed the majority of the experiments and data analysis. A.P. and L.B.B. performed RNA sequencing experiments and data analysis. L.C. helped with various experiments. Z.M.E. and D.L.R. performed bacterial cultures and species identification. T.M. performed cell-sorting experiments. J.F.P. performed histology experiments. J.D.E. and M.W.M. helped with intestinal permeability assays. H.J.G. helped with germ-free mouse experiments. N.T. performed cytokine measurements. R.B. performed RT-PCR experiments. M.B. performed 16S PCR experiments. Y.W., Y.L. and C.R.W. performed ZO-1 immunofluorescence experiments and analysis. B.D.M. helped with flow cytometry experiments. V.D., S.M.K. and M.A.Z. helped with intestinal tissue sample collection. G.C.F. performed TUNEL and caspase 3 experiments and analysis. A.M.E. performed 16S analysis. S.A.L., G.B., E.B.C., A.M.E., C.R.W., L.B., L.A.G., E.F.V. and L.B.B. provided critical intellectual input and technical support. T.M., L.A.G., L.B.B. and B.J. critically edited the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0125-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0125-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to B.J.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Mice. *Tet2*$^{-/-}$ mice have previously been described[8]. *Tet2*^{f} (B6;129S-*Tet2*$^{tm1.1Iaai}$/J) mice were crossed with the haematopoietic-specific Vav-cre line (B6.Cg-Tg(Vav1-icre)A2Kio/J), the villus and crypt small and large intestinal epithelial cell-specific Villin-cre line (B6.Cg-Tg(Vil1-cre)997Gum/J), and the myeloid cell lineage-specific LysM-cre line (Lyz2tm1(cre)Ifo)), respectively. For the experiments, >20-week-old *Tet2*$^{-/-}$ mice and littermate controls were used, unless indicated otherwise. Both female and male mice were used for experiments; no notable sex-dependent differences were found for the reported experiments. Mice were housed at the University of Chicago animal facilities under specific pathogen-free (SPF) conditions, where cages were changed on a weekly basis; ventilated cages, bedding, food and water (non-acidified) were autoclaved before use, ambient temperature maintained at 23 °C, and 5% Clidox-S was used as a disinfectant. Experimental breeding cages were randomly housed on two different racks in the vivarium, and all cages were kept on automatic 12-h light/dark cycles. Germ-free *Tet2*$^{-/-}$ mice were generated by two-stage embryo transfer, as previously described[29], bred to germ-free C57BL/6 wild-type mice to generate littermate controls and maintained in flexible film isolators in McMaster's Axenic Gnotobiotic Unit. Animal husbandry for both SPF and germ-free facilities, and experimental procedures, were performed in accordance with Public Health Service policy and approved by the University of Chicago Institutional Animal Care and Use Committees and the McMaster University Animal Care Committee.

Gnotobiotic animal husbandry. Food, bedding and water (non-acidified) were autoclaved before transfer into the sterile isolators. Cages within isolators were changed weekly, and all the cages in the vivarium were kept on 12-h light/dark cycles. Microbiology testing of faecal (experimental mice) or of caecum samples (sentinel mice; aerobic and anaerobic culture, 16S qPCR) was performed every other week to confirm germ-free status.

In vivo anti-IL-6 antibody treatment. Peripheral blood of >20-week-old *Tet2*$^{-/-}$ mice and littermate controls was analysed for myelopoiesis by flow cytometry. *Tet2*$^{-/-}$ mice selected with increased myelopoiesis (>16% CD11b^{+}Gr1^{+}) in peripheral blood, and littermate controls, were subjected to weekly intraperitoneal injections of 1 mg anti-mouse IL-6 (MP5-20F3, Bio X Cell, BE0046) or 1 mg rat IgG isotype control (HRPN, Bio X Cell, BE0088), respectively. Two *Tet2*$^{-/-}$ mice that did not respond to the anti-mouse IL-6 treatment (based on blood screening) were excluded from the study. After three weeks, mice were euthanized and analysed for signs of PMP in the peripheral blood and spleen.

In vivo Pam3CSK4 treatment. Peripheral blood of >20-week-old *Tet2*$^{f/f}$*Vav*^{cre} mice and littermate controls was analysed for myelopoiesis by flow cytometry. Mice selected with no signs of myelopoiesis (<16% CD11b^{+}Gr1^{+}) in peripheral blood and <30 g body weight were injected intraperitoneally with 100 μg TLR1/2 agonist Pam3CSK4 (InvivoGen, tlrl-pms) every four days. On day 16, mice were euthanized and analysed for signs of PMP in the peripheral blood and spleen.

DSS-induced colitis model. Peripheral blood of >20-week-old *Tet2*$^{f/f}$*Vav*^{cre} mice and littermate controls was analysed for myelopoiesis by flow cytometry. Mice selected with no signs of myelopoiesis (<16% CD11b^{+}Gr1^{+}) in peripheral blood and <30 g body weight underwent DSS colitis treatment as described[30]. In brief, mice received 3% DSS (MW: 36,000–50,000; MP Biomedicals) (w/v) in drinking water for 7 days and were then switched to regular drinking water. Weight was recorded every other day throughout the study. Mice were euthanized 30 days post-DSS feeding and analysed for PMP.

Antibiotic treatment. Mice were treated with an antibiotic cocktail as previously described[31]. All antibiotics used are listed as follows (concentration used, company, catalogue number). In detail, in the first week mice received a daily intragastric gavage with 100 μl of a mixture of kanamycin (4 mg ml^{-1}, Sigma-Aldrich, 60615), gentamicin (0.35 mg ml^{-1}, Sigma-Aldrich, G1914), colistin (8500 U ml^{-1}, Sigma-Aldrich, C4461), metronidazole (2.15 mg ml^{-1}, Sigma-Aldrich, M3761), and vancomycin (0.45 mg ml^{-1}, Sigma-Aldrich, V2002). For the following three weeks antibiotics were administered in the autoclaved (non-acidified) drinking water at 50-fold dilution except for vancomycin, which was maintained at 0.5 mg ml^{-1}. Autoclaved (non-acidified) water was used to dissolve antibiotics. Antibiotic water was prepared fresh and replaced weekly.

In vivo intestinal permeability measurement. An in vivo assay to determine intestinal epithelial barrier permeability was performed using a FITC-labelled dextran method as previously described[32]. In detail, mice were withdrawn from both food and water for 4 h, weighed and then received 60 mg per 100 g body weight of a freshly prepared FITC–dextran (MW 4,000, Sigma-Aldrich, 46944) solution diluted to 60 mg ml^{-1} in sterile PBS by oral gavage. Blood was collected by cheek bleeding (~300–500 μl) 3 h post-gavage and plasma was collected by centrifugation of blood samples at 2,000*g* for 10 min at 4 °C. Fifty microlitres of blood plasma was transferred (in duplicate) into a flat-bottom 96-well plate (Corning, 3370) and analysis of FITC–dextran concentration was performed with a fluorescence spectrophotometer setup (Synergy, BioTek) with emission and excitation wavelengths of 520 nm and 490 nm, respectively. Plasma samples were protected from light at all times. FITC–dextran concentration was determined from a standard curve generated by serial dilution of FITC–dextran.

Tissue collection and cell purification. Spleen, MLNs and bones (tibia and femur, both sides) were collected under a laminar flow hood with autoclaved tools under sterile conditions and weight of spleen was recorded. Peripheral blood was collected by cheek bleeding (after sterilizing cheeks with 70% ethanol wipes) into EDTA-coated (Thermo Fisher Scientific, MT-46034CI) 1.5-ml tubes. To obtain a single cell suspension, spleen and MLNs were mashed through a 70-μm cell strainer and bone marrow was flushed out with syringes filled with PBS and also mashed through a 70-μm cell strainer. Erylysis of 10 μl peripheral blood, spleen and bone marrow was performed using the Mouse Erythrocyte Lysing Kit (R&D Systems, WL2000). For analysis of splenic and bone marrow haematopoietic precursors the mouse Lineage Cell Depletion Kit (Miltenyi Biotec, 130-090-858) was used.

Antibodies and flow cytometry. Single cell suspensions were pelleted and resuspended in FACS buffer (PBS, 2% FCS) for immunostaining and subsequent flow cytometry analysis. Excluding the haematopoietic progenitor cell staining, all cell suspensions were incubated with Fc Block (BD, 553142) before staining with fluorophore-conjugated monoclonal antibodies. All fluorophore-conjugated antibodies used are listed as follows (clone, company, catalogue number): Gr-1 (RB6-8CA, eBioscience, 45-5931-80), c-Kit (2B8, BD Biosciences, 553356), CD11b (M1/70, eBioscience, 25-0112-82), CD34 (RAM34, eBioscience, 11-0341-85), CD16/CD32 (Fcγ RIII/II) (93, eBioscience, 25-0161-81), CD19 (1D3, BD Biosciences, 563148), CD3 (17A2, BD Biosciences, 564009), B220 (RA3-6B2, BD Biosciences, 563708), TER-119 (TER-119, BD Biosciences, 563323), CD48 (HM48-1, BD Biosciences, 561242), CD135 (A2F10.1, BD Biosciences, 562898), Ly-6C (AL-21, BD Biosciences, 560595), Sca-1 (D7, Biolegend, 108108), CD150 (TC15-12F12.2, Biolegend, 115922), CD126 (D7715A7, BD Biosciences, 740038) and Rat IgG2b (R35-38, BD Biosciences, 562603). Aqua LIVE/DEAD Fixable Aqua Dead Cell Stain Kit was purchased from Life Technologies (L34966). CD11b^{+}Gr1^{+} myeloid cells were gated on live CD45^{+} cells and lymphocytes are defined as live CD45^{+} gated CD11b^{-} cells. For phospho-flow, cells were stimulated with 0, 1 or 10 ng μl^{-1} murine IL-6 (Biolegend, 575704) for 30 min at 37 °C, 5% CO_2, fixed with Fix Buffer I (BD Biosciences, 557870) for 10 min at 37 °C, stained with Sca-1 (Biolegend, 108126) for 20 min at 4 °C and permeabilized with Perm Buffer III (BD Biosciences, 558050) and stained for phospho-Stat3 (pY705, BD Biosciences, 612569), c-Kit, CD34, CD11b, Gr1, CD19 and B220 for 30 min at room temperature. Flow cytometry analysis was performed with a nine-colour BD FACSCanto (BD Biosciences) and Aria Fusion (BD Biosciences; for cell sorting) using FlowJo software (Treestar).

RNA processing and RT–PCR. Small and large intestines were removed and transferred into cold PBS. A piece (~5 mm) of whole intestinal tissue (duodenum, jejunum, ileum and colon) was soaked in RNAlater (Qiagen, 76106) at 4 °C for 48 h and then stored at −80 °C until further analysis. For RNA extraction a Tissue-Tearor Homogenizer (Biospec) was used. RNA was prepared using the RNeasy Mini Kit (Qiagen, 74136). cDNA synthesis was performed using GoScript (Promega, A5004) according to the manufacturer's instructions. Expression analysis was performed in duplicate via RT–PCR on a Roche LightCycler 480 using SYBR Green (Clontech, 639265). Expression levels were quantified and normalized to *Gapdh* expression using the following primer pairs (all mouse):

Gapdh forward: 5′-AGGTCGGTGTGAACGGATTTG-3′; *Gapdh* reverse: 5′-TGTAGACCATGTAGTTGAGGTCA-3′; *Dsp* forward: 5′-TAC ACCTCAGGG C TGGAAAC-3′; *Dsp* reverse: 5′-GGGCCAGTCTTAGCTCCTCT-3′; *Retnlb* forward: 5′-ATGAAGCCTACACTGTGTTTCC-3′; *Retnlb* reverse: 5′-CTGCCAGAAGACGTGACACT-3′; *Tjp1* forward: 5′-ACTCCCAC TTCCCCAAAAAC-3′; *Tjp1* reverse: 5′-CCACAGCTGAAG GACTCACA -3′; *Ang4* forward: 5′-GGTTGTGATTCCTCCAACTCTG-3′; *Ang4* reverse: 5′-CTGAAGTTTTCTCCATAAGGGCT-3′; *Ocln* forward: 5′-ACTGG GTCAGGGAATATCCA-3′; *Ocln* reverse: 5′-TCAGCAGCAGCCATG TACTC-3′; *Il6* forward: 5′-CCAAGAGGTGAGTGCTTCCC-3′; *Il6* reverse: 5′-CTGTTGTTCAGACTCTCTCCCT-3′

RNA sequencing processing and data analysis. RNA sequencing libraries were prepared using the Illumina TruSeq protocol and sequenced with single-end 100-bp reads on an Illumina HiSeq2500. Adaptor sequences and low quality score bases were first trimmed using Trim Galore. The resulting reads were aligned to the mouse genome reference sequence (GRCm38/mm10) using the TopHat2 software package[33] with a TopHat transcript index from Ensembl. The number of read fragments overlapping with annotated exons of genes was tabulated using HTSeq[34] using the following parameters: -q -m intersection-nonempty -s no. Non-coding or low-expressed genes with an average read count lower than 10 were discarded, resulting in 15,045 genes in total. Using normalized gene counts for three *Tet2*$^{+/+}$ and three *Tet2*$^{-/-}$ samples for each of the four tissue compartments, we identified differentially expressed genes using the R package limma[35]. Our linear model had the following design: Expression ~ Genotype:Tissue. Once we identified jejunum as the tissue with the largest differences in gene expression between *Tet2*$^{+/+}$ and

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Tet2$^{-/-}$ samples, we increased our sample size for this tissue for a total of seven *Tet2*$^{+/+}$ and seven *Tet2*$^{-/-}$ jejunum samples. We corrected for batch effects between the first and second set of samples using ComBAT[36]. We considered a gene as differentially expressed if statistically supported at a *Q*-value[37] false discovery rate <0.1 and a $|\log_2(\text{fold change})| > 0.5$.

Bacterial cultures and species identification. Organs were aseptically removed, weighed and transferred immediately to an anaerobic chamber (Coy Systems) for manual homogenization in 300 µl of pre-reduced (autoclaved, sterile filtered (0.22 µm), oxygen reduced with vacuum degasification) PBS + 0.1% L-cysteine (Sigma-Aldrich, 168149). Homogenates (100 µl per plate) were plated onto BD BBL Brucella agar with 5% sheep blood, Hemin and vitamin K1 agar plates (Fisher Scientific, L97848) and incubated anaerobically. For aerobic cultures tissue homogenates were plated onto Teknova brain heart infusion (BHI) agar plates (Fisher Scientific, 50-841-098) and incubated aerobically. Plates were sealed with parafilm and incubated upside down and after 48 h (aerobic) or 120 h (anaerobic) the colony-forming units (CFUs) were quantified. Results are displayed as CFUs per gram (g) tissue.

Representative bacterial colonies grown on agar plates were picked with sterile pipette tips and stored at −80 °C until analysis. At day of analysis picked bacterial colonies were thawed at room temperature, resuspended with 6 µl of sterile water and bacteria were lysed by heating up the samples to 95 °C for 10 min. Samples were subsequently cooled down to 4 °C and then the DNA (2 µl) was used as template DNA in PCR reactions amplifying the 16S rRNA gene using universal bacterial 16S rRNA primers (27F, 5′-AGAGTTTGATCMTGGCTCAG-3′ and 1525R, 5′-AAGGAGGTGATCCAGCC-3′) with reaction conditions: 95 °C for 5 min followed by 35 cycles of 95 °C for 30 s, 55 °C for 30 s, 72 °C for 2 min and then 72 °C for 20 min. The amplification product (10 µl) was incubated with 2 µl ExoSAP-IT (Thermo Fisher, 78200.200.UL) for 37 °C for 15 min, followed by 80 °C for 15 min. Amplicons were sequenced by capillary sequencing, and the resulting sequences were analysed using BLASTN and the 16S ribosomal RNA sequences database for species identification[38].

Bacterial lysates. Individual bacterial strains were grown up anaerobically in lactobacillus MRS broth (Fisher Scientific, DF0881-17-5) for 48 h, and sonicated for 3 × 20 s in cold PBS. The homogenate was centrifuged at 10,000*g* for 20 min to remove debris and filtered through a 0.2-µM filter. 16S copies of the homogenate were quantified by qPCR using the universal primers (340F, 5′-ACTCCTACGGGAGGCAGCAGT-3′) and (514R, 5′-ATTACCGCGGCTGCTGGC-3′) and purified genomic DNA from *Blautia producta* (Prevot) (ATCC, 27340D-5) as a standard[39].

HEK-TLR2 Quanti-Blue secreted embryonic alkaline phosphatase reporter assay. HEK-Blue TLR2 cells (provided by the Gajewski laboratory) were plated in HEK-Blue detection medium (InvivoGen, hb-det2) containing a substrate for alkaline phosphatase at 5×10^4 cells per well in 96-well plates and incubated with different concentrations of Pam3CSK4 (InvivoGen, tlrl-pms), LPS (InvivoGen, tlrl-peklps) or bacterial lysates normalized to 16S rRNA copies. Alkaline phosphatase activity was measured after 5 h by reading optical density at 620 nm.

DNA extraction from tissue or faeces for 16S qPCR. Total DNA was extracted from organs (collected under sterile conditions, ~30 mg spleen, ~30 mg MLN and 50 µl peripheral blood) and faeces using the DNeasy Blood & Tissue Kit (Qiagen, 69504) and the Fast DNA Stool Mini Kit (Qiagen, 51604), respectively. Quantitative PCR (qPCR) of 16S rRNA-encoding genes was performed as previously described[30]. In brief, qPCR was performed on a Roche LightCycler 480 (Roche Scientific) using the primers (340F, 5′-ACTCCTACGGGAGGCAGCAGT-3′) and (514R, 5′-ATTACCGCGGCTGCTGGC-3′). Reactions were run at 95 °C for 3 min, followed by 40 cycles of 95 °C for 15 min and 63 °C for 60 s. Specific amplification of targets was quantified using dilution curves of a purified pCR4-TOPO vector (Invitrogen) containing a cloned 16S rRNA-encoding gene from *Blautia producta* (Prevot) (ATCC, 27340D-5) as a standard[39] for both blood and faeces. Standards ranging in concentration from 10^8–10^0 plasmid copies per µl were run in parallel with our samples for both blood and faeces during each qPCR run. Using those results, a standard curve was generated to quantify the number of copy numbers within the samples. To determine the bacterial load in the faeces samples, the results were normalized to faecal weight. Bacterial 16S rRNA gene qPCRs from MLN and spleen were normalized to the host murine *Ifnb1* gene (forward: 5′-CCATCCAAGAGATGCTCCAG-3′; reverse 5′-GTGGAGAGCAGTTGAGGACA-3′). C57BL/6 wild-type germ-free mice served as negative control (Fig. 1b).

16S rRNA amplicon library preparation, sequencing and data analysis. Faeces, jejunal and colon contents, and mucosal scrapings from the jejunum were collected and snap-frozen in liquid nitrogen. Bacterial DNA was extracted using the Fast DNA Stool Mini Kit (Qiagen, 51604). 16S rRNA amplicon library preparation and sequencing was performed as previously described[30]. Raw sequencing data were de-multiplexed, and partially overlapping paired-end reads were merged using illumina-utils[40]. Mismatches at the overlapping regions of pairs were resolved using the base with the higher *Q*-score, and the merged sequences were kept for downstream analyses only (1) if they contained less than three mismatches at the overlapping region, and (2) if 66% of the bases in the first half of each read had an average *Q*-score of 30. The quality-filtered reads were partitioned into ecologically relevant units using minimum entropy decomposition (MED)[41] with default parameters. Using Shannon entropy, MED resolves a given amplicon dataset iteratively into oligotypes, which are able to describe differences between microbial groups at a single-nucleotide resolution[22]. The taxonomy of oligotypes was inferred through GAST[42]. The metaMDS function in the vegan[43] library was used to generate ordinations of samples based on their oligotype profiles using non-metric multidimensional scaling with Bray–Curtis distances, and the envfit function in vegan with 999 permutations was used to test whether the genotype or the source of samples were significantly associated with differences between microbiomes. The pheatmap library for R was used to generate heat-map displays of the sample oligotype profiles, and the hierarchical clustering of samples in heat maps employed Bray–Curtis distances, and the average linkage algorithm.

Analysis of cytokine production in blood plasma and intestinal explants. Blood was collected by cheek bleeding (after sterilizing cheeks with 70% ethanol wipes) in EDTA (Thermo Fisher Scientific, MT-46034CI)-coated tubes, centrifuged at 2,000*g* for 10 min at 4 °C and plasma was stored at −80 °C until cytokine analysis. One-centimetre pieces from the jejunum and colon were collected, opened longitudinally, washed in PBS and explants were cultured in 500 µl RPMI 1640 medium (Fisher Scientific, MT-10-040-CV) containing 10% FBS, 200 U ml^{-1} penicillin, and 200 µg ml^{-1} streptomycin (MP Biomedicals, 091670249) for 48 h. The IL-6 cytokine amount in blood plasma and in supernatants of intestinal explants was determined with BioPlex multianalyte technology (Biorad) according to manufacturer's instructions.

In vitro haematopoietic methylcellulose colony-forming assay. Similar to a related study[9], lineage depleted (mouse Lineage Cell Depletion Kit (Miltenyi Biotec, 130-090-858) haematopoietic progenitors of the bone marrow and the spleen were seeded in duplicate in cytokine-supplemented methylcellulose (MethoCult, M3434, STEMCELL Technologies) at a density of 5,000 cells per plate (bone marrow) or 25,000 cells per plate (spleen). Colonies were counted every 7 days and 10,000 cells per plate were seeded for replatings. The average number of colonies per 10,000 plated cells is shown for 5 sequential platings. For anti-mouse IL-6 antibody (MP5-20F3, Bio X Cell, BE0046) and isotype (HRPN, Bio X Cell, BE0088) treatment, anti-IL-6 antibody or isotype was freshly added to the cytokine-supplemented methylcellulose medium before every replating at a concentration of 100 µg ml^{-1}.

Haematoxylin and eosin staining. Jejunum was flushed with cold PBS (GE Healthcare Life Science, SH30028.02) followed by immediate flushing with 10% formalin (Fisher Scientific, L14416). A section of ~3 cm from jejunum was cut with a razor blade, cut open longitudinally and pinned out with needles in wax boxes. Tissue was fixed for 24 h at room temperature and pre-embedded in 2% agar (Sigma, A7921). Haematoxylin and eosin staining was performed on 5-µm paraffin-embedded intestinal sections. The haematoxylin and eosin stained slides were digitized (objective: ×20; camera: CIS 3CCD 2 Mega Pixel) with the Pannoramic MIDI Scan System (3DHISTECH). The Panoramic viewer software (3DHISTECH) was used for imaging.

Immunofluorescence staining. For ZO-1 staining, jejunum and proximal colon was dissected and cross-sections were frozen in optimum cutting temperature compound (Tissue-Tek). Frozen sections of 5-µm thicknesses were cut and fixed in 1% paraformaldehyde in 1× PBS for 30 min at room temperature. After washing in PBS, three 5-min incubations in PBS with 50 mM NH_4Cl were performed, followed by permeabilization in PBS with 0.5% NP-40 for 20 min, and three additional 5-min washes in PBS. Tissue sections were then incubated with mouse IgG ZO-1 monoclonal antibody (ZO1-1A12, Thermo Fisher Scientific, 33-9100) in PBS with 10% normal donkey serum overnight at 4 °C, washed five times with PBS with 1% normal donkey serum (5 min each) and incubated with Alexa Fluor 594-conjugated donkey anti-mouse (IgG, Jackson ImmunoResearch, 715-586-151) and Hoechst 33342 (ThermoFisher Scientific, H3570) for 1 h at room temperature. After incubation, tissue was washed five times with PBS with 1% normal donkey serum (5 min each) and mounted in Pro-Long Gold slides followed by imaging using an Olympus IX81 inverted microscope with a Photometrics CoolSNAP HQ2 camera. For cleaved caspase 3 and TUNEL staining, jejunum was dissected and cross-sections were fixed in 10% formalin, and then processed for paraffin sectioning. Five-micrometre sections were dewaxed by immersion in xylene and hydrated by serial immersion in ethanol and PBS. Antigen retrieval was performed by incubating sections in a pressure cooker (Cuisinart) for 15 min in Target Retrieval Solution (DAKO, S1699). Sections were washed with PBS (twice for 10 min), and blocking buffer (TBS containing 10% BSA (EMD Millipore Sigma, 82-045-1) and OmniPur Triton X-100 (0.3%, EMD Millipore Sigma, 9400) was added for 1 h. Sections were incubated with anti-pan-keratin (dilution 1:200, ABCAM, ab6401) and anti-cleaved caspase 3 (dilution 1:500, Cell Signaling Technology, 9661S) antibody in blocking buffer overnight at 4 °C and then incubated with goat

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

anti-mouse IgG1 Alexa Fluor 488 (Thermo Fisher Scientific, A21121) or goat anti-rabbit Alexa Fluor 594 (Thermo Fisher Scientific, A11037) labelled secondary antibodies (dilution 1:400, Thermo Fisher Scientific) for 30 min. Sections were mounted with Fluoromount-G (Beckman Coulter, 731604). DNA fragmentation in gut sections was assayed by TUNEL assay in formalin-fixed sections according to the manufacturer's instructions (Roche Applied Science, 12156792910). Sections were stained with DAPI (10 μg ml^{-1}, Invitrogen, D1306) and mounted with Fluoromount-G (Beckman Coulter, 731604). As a positive control, jejunum sections of VDTR mice (mice expressing diphtheria toxin receptor under the villin promoter (epithelial specific)) treated with 10 ng of diphtheria toxin per g of body weight, for 12 h, were used[44].

Statistical analysis. The majority of experiments were repeated at least three times to obtain data for indicated statistical analyses. Mice were allocated to experimental groups on the basis of their genotype and randomized within the given sex- and age-matched group. Because our mice were inbred and matched for age and sex, we always assumed similar variance between the different experimental groups. We did not perform an a priori sample size estimation but always used as many mice per group as possible in an attempt to minimize type I and type II errors. Investigators were not blinded during experiments and outcome assessment, except for microscopic analysis of fluorescent immunostaining, which was performed blinded. For bacterial culture experiments (aerobic and anaerobic) (Fig. 1d) a representative image out of five biological replicate experiments is shown. For microscopy analyses (Extended Data Figs. 2a–c, 4d, 10g) representative images out of at least three biological replicate experiments are shown. All experimental and control animals were littermates and none were excluded from the analysis at the time of collection. The number of mice per group is described in the corresponding figure legends as n and all quantitative data are presented as mean ± error of the mean (s.e.m.), unless otherwise indicated. Data were first analysed for normal distribution using D'Agostino and Pearson omnibus normality tests. Normally distributed data were analysed using a paired or unpaired two-tailed Student's t-test for single comparisons, and one-way or two-way ANOVA for multiple comparisons. ANOVA analysis was followed by a Sidak's post hoc test. Data that were not normally distributed were analysed using unpaired two-tailed Mann–Whitney U-test for single comparisons, and Kruskal–Wallis, Dunn's post hoc test for multiple comparisons. Correlations were calculated using the Pearson and Spearman correlation. Figures and statistical analysis were generated using GraphPad Prism 6 (GraphPad Software). The statistical test used and P values are indicated in each figure legend. P values of <0.05 were considered to be statistically significant. $*P<0.05$, $**P<0.01$, $***P<0.001$ and $****P<0.0001$.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

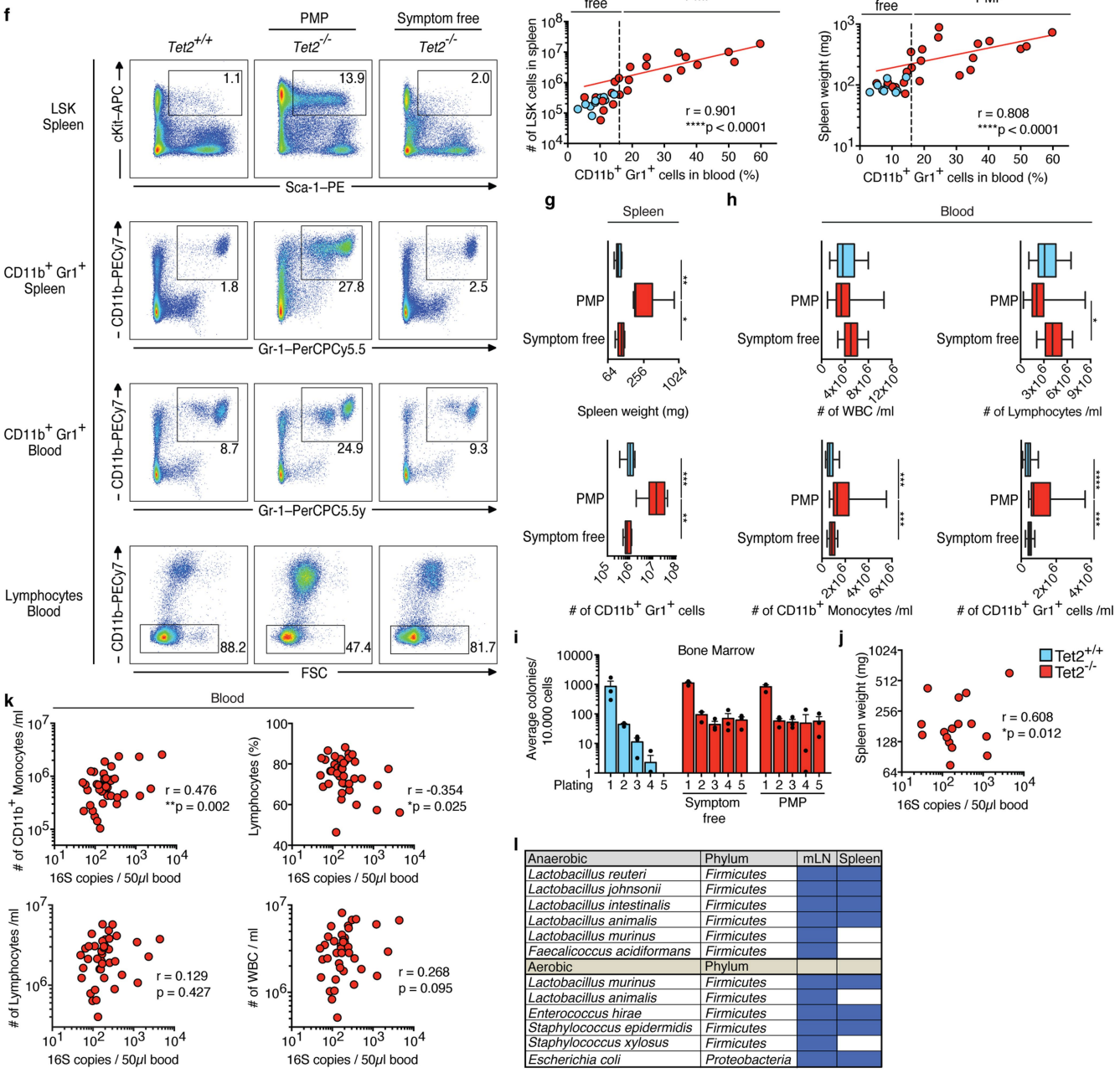
Data availability. Datasets that support the findings of this study have been deposited online in the Gene Expression Omnibus (GEO) under accession number GSE99333. Source Data for all mouse experiments have been provided. All other data are available from the corresponding author upon reasonable request.

29. Slack, E. et al. Innate and adaptive immunity cooperate flexibly to maintain host–microbiota mutualism. *Science* **325**, 617–620 (2009).
30. Meisel, M. et al. Interleukin-15 promotes intestinal dysbiosis with butyrate deficiency associated with increased susceptibility to colitis. *ISME J.* **11**, 15–30 (2017).
31. Stefka, A. T. et al. Commensal bacteria protect against food allergen sensitization. *Proc. Natl Acad. Sci. USA* **111**, 13145–13150 (2014).
32. Furuta, G. T. et al. Hypoxia-inducible factor 1-dependent induction of intestinal trefoil factor protects barrier function during hypoxia. *J. Exp. Med.* **193**, 1027–1034 (2001).
33. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
34. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
35. Ritchie, M. E. et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
36. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
37. Storey, J. D., Bass, A. J., Dabney, A. & Robinson, D. qvalue: Q-value estimation for false discovery rate control. R package version 2.8.0. http://github.com/jdstorey/qvalue (2015).
38. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
39. Huse, S. M. et al. Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects. *Microbiome* **2**, 5 (2014).
40. Eren, A. M., Vineis, J. H., Morrison, H. G. & Sogin, M. L. A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS ONE* **8**, e66643 (2013).
41. Eren, A. M. et al. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* **9**, 968–979 (2015).
42. Huse, S. M. et al. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* **4**, e1000255 (2008).
43. Oksanen, J., Kindt, R., Legendre, P., O'Hara, B. & Stevens, M. H. Vegan: community ecology package. https://CRAN.R-project.org/package=vegan (2007).
44. Cummings, R. J. et al. Different tissue phagocytes sample apoptotic cells to direct distinct homeostasis programs. *Nature* **539**, 565–569 (2016).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

a

Disease phenotype	*Tet2*$^{+/+}$	*Tet2*$^{+/-}$	*Tet2*$^{-/-}$
Splenomegaly (> 150mg)	0/10 (0%)	4/10 (40%)	9/14 (64%)
Extramedullary hematopoiesis (splenic LSK cells; > 5x10^5)	0/10 (0%)	6/10 (60%)	11/14 (79%)
Splenic myeloproliferation (CD11b$^+$ Gr1$^+$; > 1.6x10^6)	0/10 (0%)	4/10 (40%)	10/14 (71%)
PB myeloproliferation (CD11b$^+$ Gr1$^+$; > 5x10^5)	0/10 (0%)	4/10 (40%)	6/14 (43%)



l

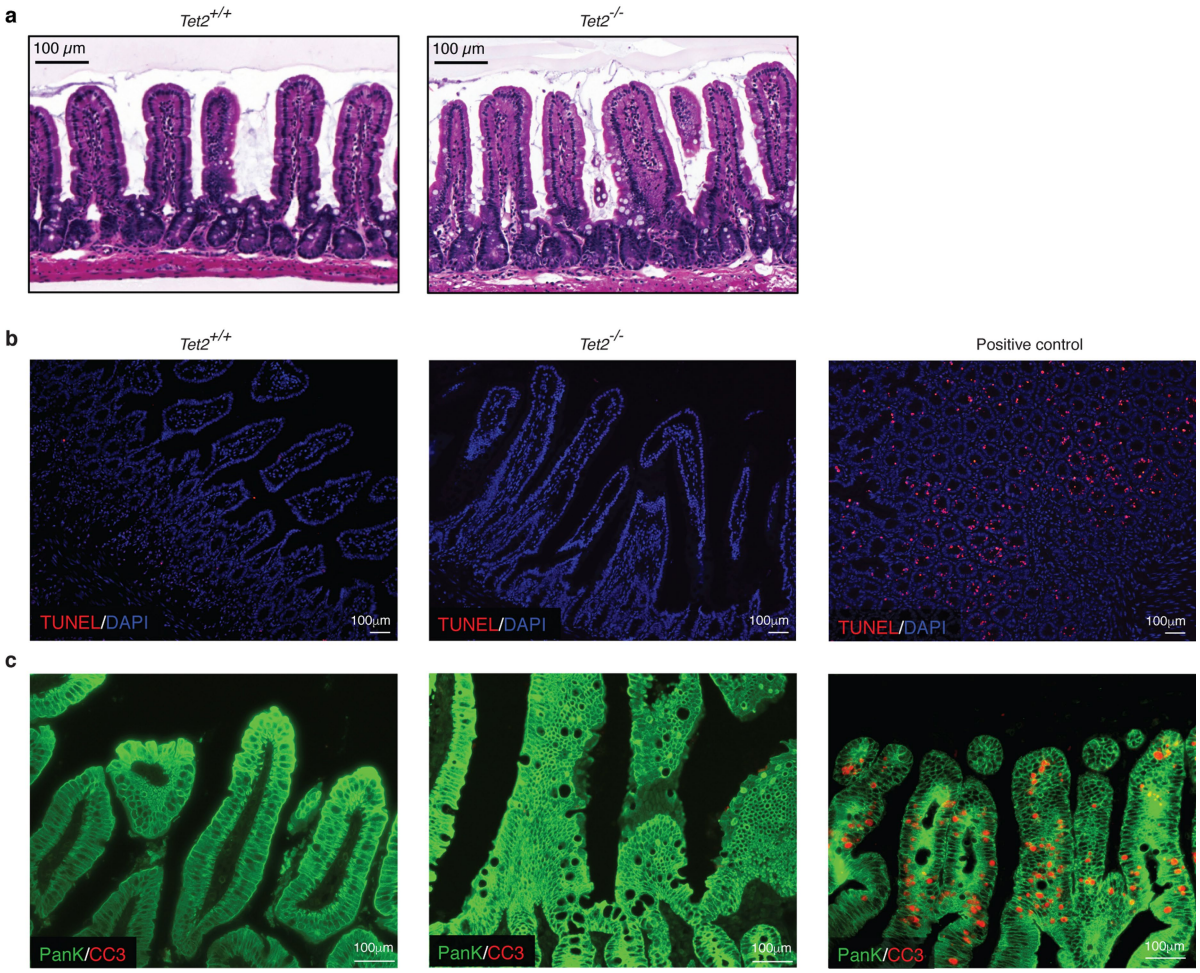
Anaerobic	Phylum	mLN	Spleen
Lactobacillus reuteri	*Firmicutes*	■	■
Lactobacillus johnsonii	*Firmicutes*	■	■
Lactobacillus intestinalis	*Firmicutes*	■	■
Lactobacillus animalis	*Firmicutes*	■	■
Lactobacillus murinus	*Firmicutes*	■	
Faecalicoccus acidiformans	*Firmicutes*	■	
Aerobic	Phylum		
Lactobacillus murinus	*Firmicutes*	■	■
Lactobacillus animalis	*Firmicutes*	■	
Enterococcus hirae	*Firmicutes*	■	■
Staphylococcus epidermidis	*Firmicutes*	■	■
Staphylococcus xylosus	*Firmicutes*	■	
Escherichia coli	*Proteobacteria*	■	■

Extended Data Fig. 1 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

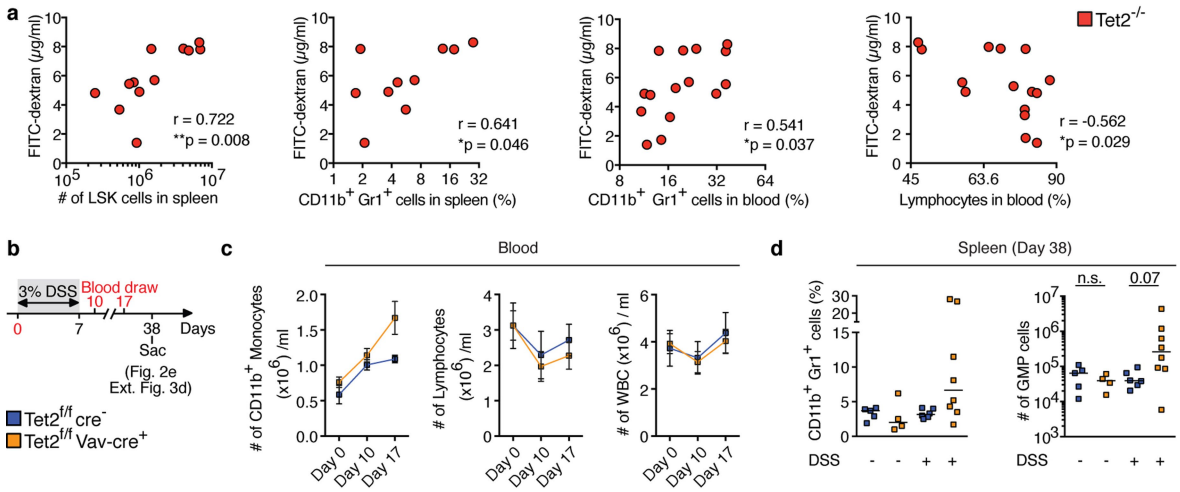
Extended Data Fig. 1 | *Tet2*$^{-/-}$ mice display variability in PMP penetrance that correlates with bacteraemia. a, Penetrance of PMP assessed by splenomegaly, splenic LSK cell expansion and CD11b^{+}Gr1^{+} myeloid cell expansion in spleen and peripheral blood. **b**–**e**, Correlation between frequency of peripheral-blood CD11b^{+}Gr1^{+} myeloid cells (peripheral-blood biomarker) and numbers of CD11b^{+}Gr1^{+} myeloid cells in spleen (**b**) and peripheral blood (**c**), numbers of LSK cells (**d**) and splenomegaly (**e**), in *Tet2*$^{+/+}$ and *Tet2*$^{-/-}$ mice. In **b**–**e**, dotted vertical lines indicate level of 16% for peripheral-blood CD11b^{+}Gr1^{+} myeloid cells (level determined by the distribution in wild-type littermates). The peripheral-blood biomarker (frequency of CD11b^{+}Gr1^{+} myeloid cells) was used to categorize *Tet2*$^{-/-}$ mice either as symptom free ($\leq$16% CD11b^{+}Gr1^{+} myeloid cells) or as mice with PMP (>16% CD11b^{+}Gr1^{+} myeloid cells). Spearman correlation test excluding *Tet2*$^{+/+}$ mice. Red lines indicate regression line calculated for *Tet2*$^{-/-}$ data points ($n = 10$ (blue), 23 mice (red). **f**–**i**, Symptom-free, *Tet2*$^{-/-}$ mice with PMP and littermate controls were used. **f**, Representative dot blots are shown of LSK cells in the spleen, CD11b^{+}Gr1^{+} myeloid cells in the spleen and peripheral blood, and lymphocytes in peripheral blood. Data are representative of five independent experiments with similar results. **g**, Spleen weight (top; $n = 11$ (blue), 9 (red, PMP) or 5 mice (red, symptom free)) and numbers of CD11b^{+}Gr1^{+} myeloid cells in the spleen (bottom; $n = 10$ (blue), 12 (red, PMP) or 7 mice (red, symptom free)). **h**, Numbers per ml of peripheral blood of live CD45^{+} gated white blood cells (WBC) (top left), lymphocytes (top right), CD11b^{+} monocytes (bottom left) and CD11b^{+}Gr1^{+} myeloid cells (bottom right) ($n = 15$ (blue), 15 (red, PMP) or 32 mice (red, symptom free)). **i**, In vitro HSC self-renewal colony-forming assay of haematopoietic progenitors of the bone marrow ($n = 3$ mice). Mean ± s.e.m. **j**, **k**, Correlation between 16S gene copies in the peripheral blood and spleen weight ($n = 16$ mice) (**j**) or numbers of CD11b^{+} monocytes (**k** top left), percentage of lymphocytes (**k** top right), numbers of lymphocytes (**k** bottom left), and numbers of leukocytes (WBC) (**k** bottom right) in the peripheral blood ($n = 40$ mice). Pearson correlation test. **l**, Bacterial colonies from Fig. 1e, f identified by 16S sequencing; blue rectangles indicate presence of bacteria. In **g**, **h**, boxes represent median values and interquartile ranges; whiskers represent minimum and maximum values. One-way ANOVA, Sidak's post hoc test. Data are representative of at least three independent experiments. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



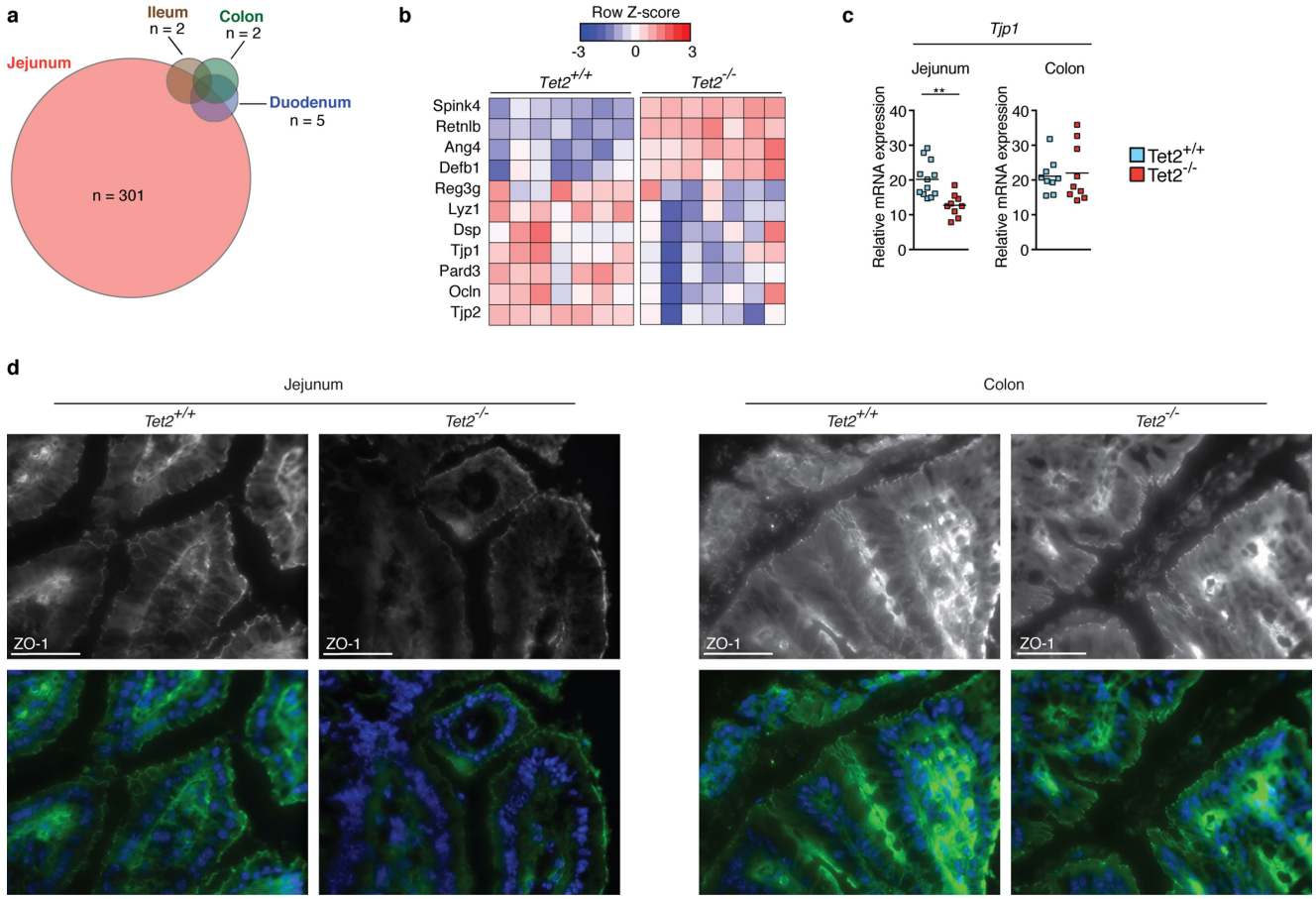
Extended Data Fig. 2 | *Tet2*$^{-/-}$ mice display no anatomical changes in the jejunum. a, Representative images of haematoxylin and eosin-stained sections of the jejunum. **b**, **c**, Immunofluorescence analyses of sections of the jejunum of *Tet2*$^{+/+}$ and *Tet2*$^{-/-}$ mice. **b**, No difference was observed between the groups in the number of apoptotic cells using the TUNEL assay (red). Nuclei of intestinal cells were visualized with DAPI (blue). **c**, No difference was observed in the number of apoptotic (cleaved caspase 3-positive (CC3), red) epithelial cells (pan-keratin-positive, (PanK), green) between the groups. In **b**, **c**, Jejunum sections of diphtheria-toxin-treated VDTR mice were used as a positive control. Data are representative of at least two independent experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



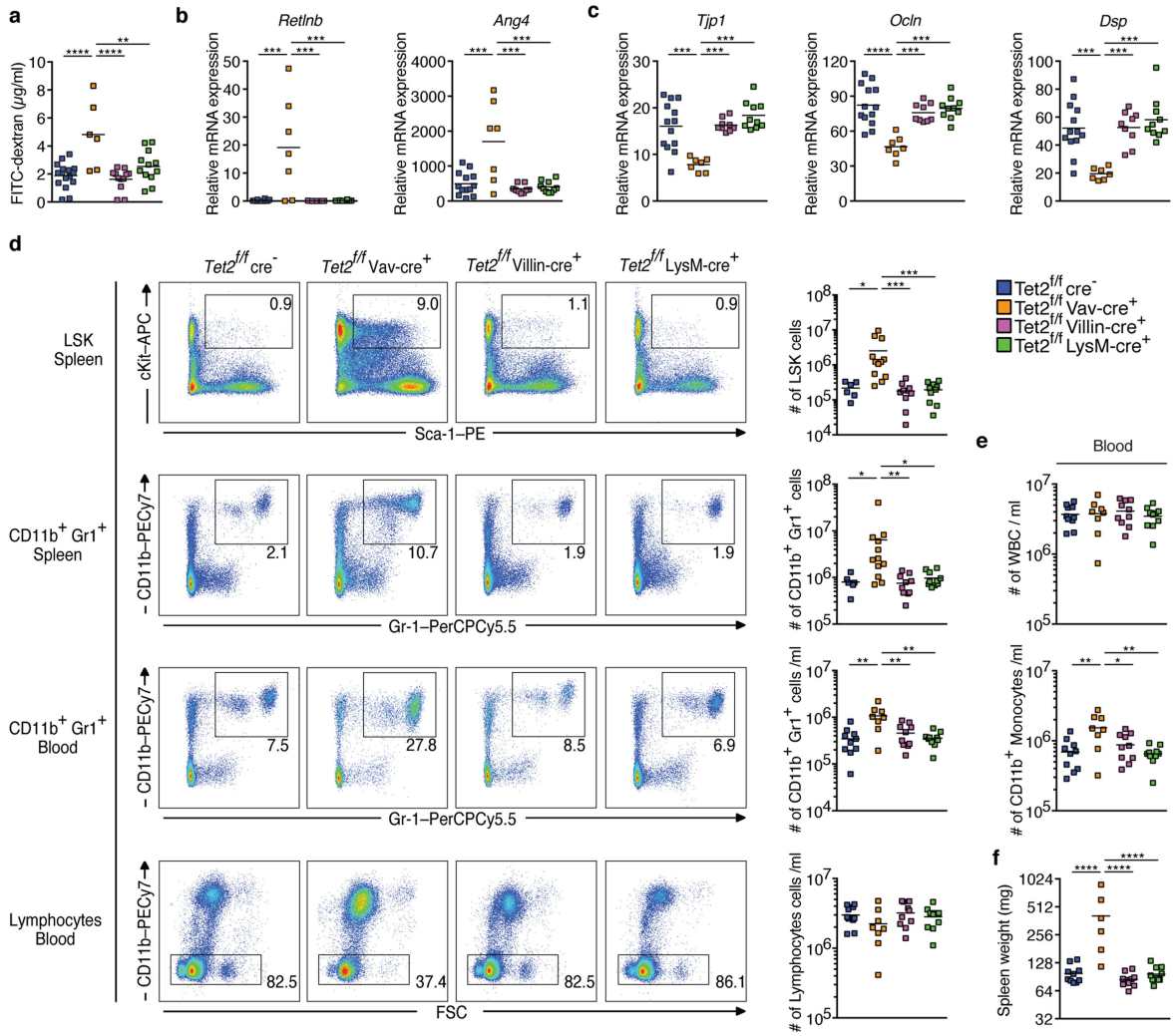
Extended Data Fig. 3 | Disruption of barrier integrity drives PMP. **a**, FITC–dextran concentration in blood plasma correlates with numbers of LSK cells (left, $n = 12$ mice), frequency of $CD11b^+Gr1^+$ myeloid cells in the spleen (middle left, $n = 10$ mice), peripheral blood (middle right, $n = 15$ mice) and frequency of lymphocytes (right, $n = 15$ mice). Pearson correlation test. **b**, Schematic of DSS treatment of symptom-free *Tet2*$^{f/f}$*Vav*^{cre} mice that are over 20 weeks old, and littermate controls. **c**, Numbers of $CD11b^+$ monocytes (left), lymphocytes (middle) and leukocytes (WBC) (right) before, during and after DSS treatment ($n = 5$ (blue) or 8 (orange) mice). Mean ± s.e.m. **d**, Percentage of $CD11b^+Gr1^+$ myeloid cells (left) and numbers of GMP (right) at end point analysis ($n = 5$ (*Tet2*$^{f/f}$*cre*$^-$, without DSS), 4 (*Tet2*$^{f/f}$*Vav*^{cre}, without DSS), 6 (*Tet2*$^{f/f}$*cre*$^-$, with DSS) or 8 (*Tet2*$^{f/f}$*Vav*^{cre}, with DSS) mice). Centre is median, Kruskal–Wallis, Dunn's post hoc test. $^*P < 0.05$, $^{**}P < 0.01$. Data are representative of at least two independent experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



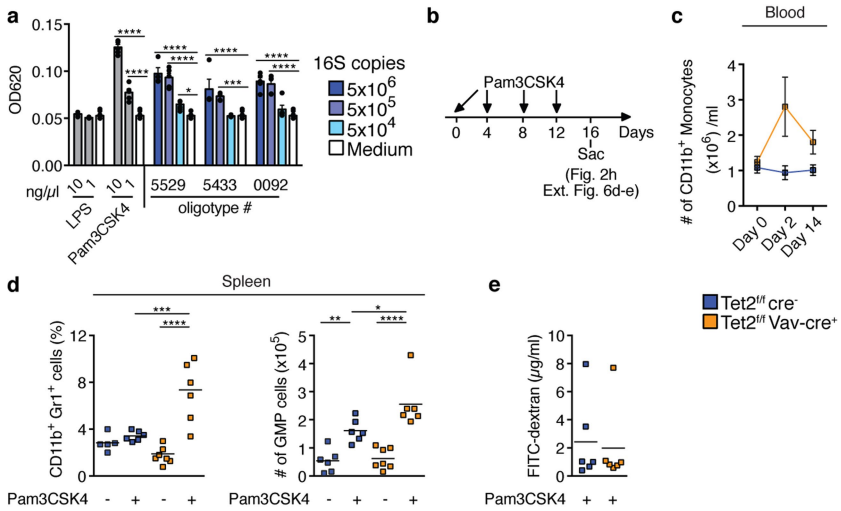
Extended Data Fig. 4 | Tight junction ZO-1 is markedly reduced in the jejunum but not in the colon of *Tet2*$^{-/-}$ mice. a, RNA sequencing of whole intestinal tissue of duodenum, jejunum, ileum and colon of three *Tet2*$^{-/-}$ and three *Tet2*$^{+/+}$ mice. Venn diagram illustrating the number of differentially expressed genes ($|\log_2(\text{fold change})| > 0.5$ and false discovery rate < 0.1); see Supplementary Table in *Tet2*$^{-/-}$ mice compared to littermate controls. **b**, RNA sequencing of the jejunum of seven *Tet2*$^{-/-}$ and seven *Tet2*$^{+/+}$ mice (the full list of differentially expressed genes is shown in Supplementary Table. Heat map of a selection of antimicrobial peptides and tight junction genes. **c**, Gene expression of tight junction gene (ZO-1, *Tjp1*) in the jejunum (left; $n = 12$ (*Tet2*$^{+/+}$) or 9 (*Tet2*$^{-/-}$) mice) and colon (right; $n = 9$ mice in both cases). Centre is mean, two-tailed unpaired *t*-test. **d**, ZO-1 immunofluorescence shows reduced tight junction staining of epithelial cells in the jejunum (left) but no difference in colonic epithelial cells (right) of *Tet2*$^{-/-}$ mice. Representative images are shown from *Tet2*$^{+/+}$ ($n = 3$ mice) and *Tet2*$^{-/-}$ mice ($n = 7$ mice). Scale bars, 100 μm. Green, ZO-1; blue, DAPI. $^{**}P < 0.01$. Data are representative of at least two independent experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



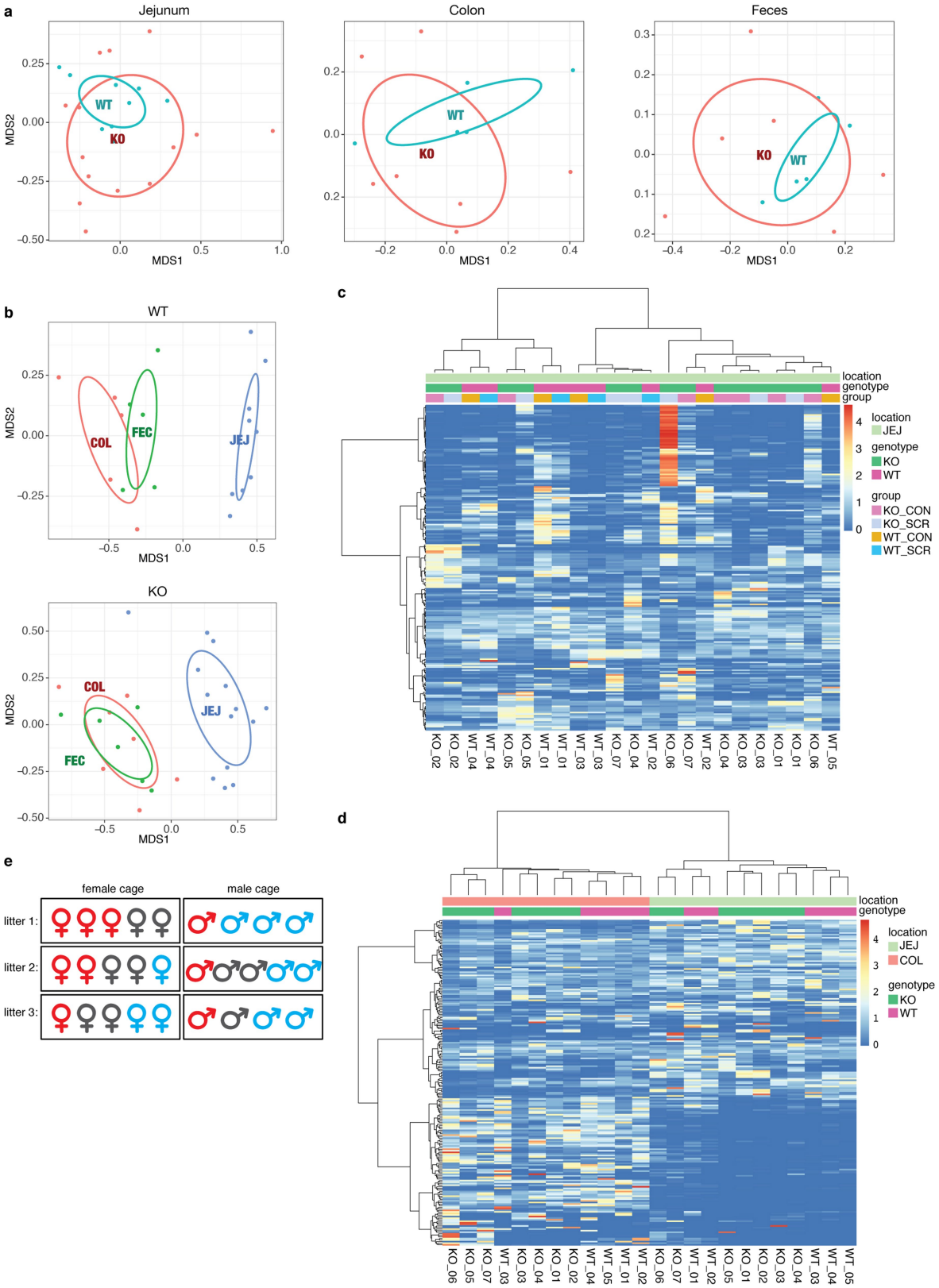
Extended Data Fig. 5 | Absence of *Tet2* expression in haematopoietic cells leads to barrier dysfunction and PMP. a, Intestinal permeability measurement by blood plasma FITC–dextran concentrations ($n = 16$ (*Tet2*$^{f/f}$*cre*$^{-}$), 6 (*Tet2*$^{f/f}$*Vav*^{cre}), 12 (*Tet2*$^{f/f}$*Villin*^{cre}) or 12 (*Tet2*$^{f/f}$*LysM*^{cre}) mice). **b**, **c**, Gene expression of antimicrobial peptides *Retlnb* (**b** left), *Ang4* (**b** right) and (**c**) tight junction genes *ZO-1* (*Tjp1*) (left), occludin (*Ocln*) (middle), and desmoplakin (*Dsp*) (right) in the jejunum ($n = 13$ (*Tet2*$^{f/f}$*cre*$^{-}$), 7 (*Tet2*$^{f/f}$*Vav*^{cre}), 9 (*Tet2*$^{f/f}$*Villin*^{cre}) or 10 (*Tet2*$^{f/f}$*LysM*^{cre}) mice). **d**, Representative dot blots (left) and numbers of LSK cells (right) in the spleen, CD11b^{+}Gr1^{+} myeloid cells in the spleen ($n = 6$ (*Tet2*$^{f/f}$*cre*$^{-}$), 12 (*Tet2*$^{f/f}$*Vav*^{cre}), 10 (*Tet2*$^{f/f}$*Villin*^{cre}) or 10 (*Tet2*$^{f/f}$*LysM*^{cre}) mice) and peripheral blood, and lymphocytes in the peripheral blood ($n = 10$ (*Tet2*$^{f/f}$*cre*$^{-}$), 8 (*Tet2*$^{f/f}$*Vav*^{cre}), 10 (*Tet2*$^{f/f}$*Villin*^{cre}) or 10 (*Tet2*$^{f/f}$*LysM*^{cre}) mice). **e**, White blood cell (WBC) count (top) and cell number per ml of blood of CD11b^{+} monocytes (bottom) ($n = 10$ (*Tet2*$^{f/f}$*cre*$^{-}$), 8 (*Tet2*$^{f/f}$*Vav*^{cre}), 10 (*Tet2*$^{f/f}$*Villin*^{cre}) or 10(*Tet2*$^{f/f}$*LysM*^{cre}) mice). **f**, Spleen weight is shown ($n = 10$ (*Tet2*$^{f/f}$*cre*$^{-}$), 6 (*Tet2*$^{f/f}$*Vav*^{cre}), 10 (*Tet2*$^{f/f}$*Villin*^{cre}) or 10 (*Tet2*$^{f/f}$*LysM*^{cre}) mice). In **a**–**f**, centre is mean, one-way ANOVA, Sidak's post hoc test. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. Data are representative of three independent experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Microbial signals are sufficient for PMP in *Tet2*$^{-/-}$ mice. a, TLR2 activation measured using a HEK-TLR2 reporter assay ($n = 5$ biological replicates). **b**, Schematic of Pam3CSK4 in vivo treatment of symptom-free *Tet2*$^{f/f}$*Vav*^{cre} mice that are over 20 weeks old, and littermate controls. **c**, Numbers of CD11b^{+} monocytes prior (day 0), during treatment (day 2) and at end point analysis (day 14) ($n = 6$ mice). Mean ± s.e.m. **d**, Percentage of CD11b^{+}Gr1^{+} myeloid cells (left, $n = 5$ (*Tet2*$^{f/f}$*cre*$^{-}$, no Pam3CSK4), 6 (*Tet2*$^{f/f}$*cre*$^{-}$, with Pam3CSK4), 7 (*Tet2*$^{f/f}$*Vav*^{cre}, no Pam3CSK4) or 6 (*Tet2*$^{f/f}$*Vav*^{cre}, with Pam3CSK4) mice) and numbers of GMP cells (right, $n = 6$ (*Tet2*$^{f/f}$*cre*$^{-}$, no Pam3CSK4), 6 (*Tet2*$^{f/f}$*cre*$^{-}$, with Pam3CSK4), 7 (*Tet2*$^{f/f}$*Vav*^{cre}, no Pam3CSK4) or 6 (*Tet2*$^{f/f}$*Vav*^{cre}, with Pam3CSK4) mice). **e**, Intestinal permeability measurement by blood plasma FITC–dextran concentrations at end point analysis ($n = 6$ mice in all cases). **a**, **c**–**e**, Centre is mean. **a**, **d**, One-way ANOVA, Sidak's post hoc test, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. Data are representative of at least two independent experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

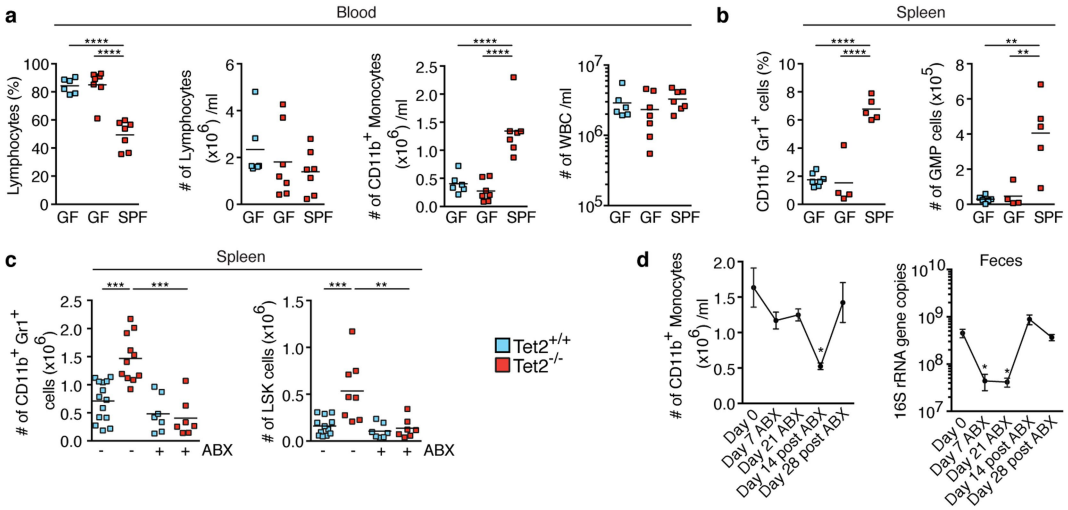


Extended Data Fig. 7 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

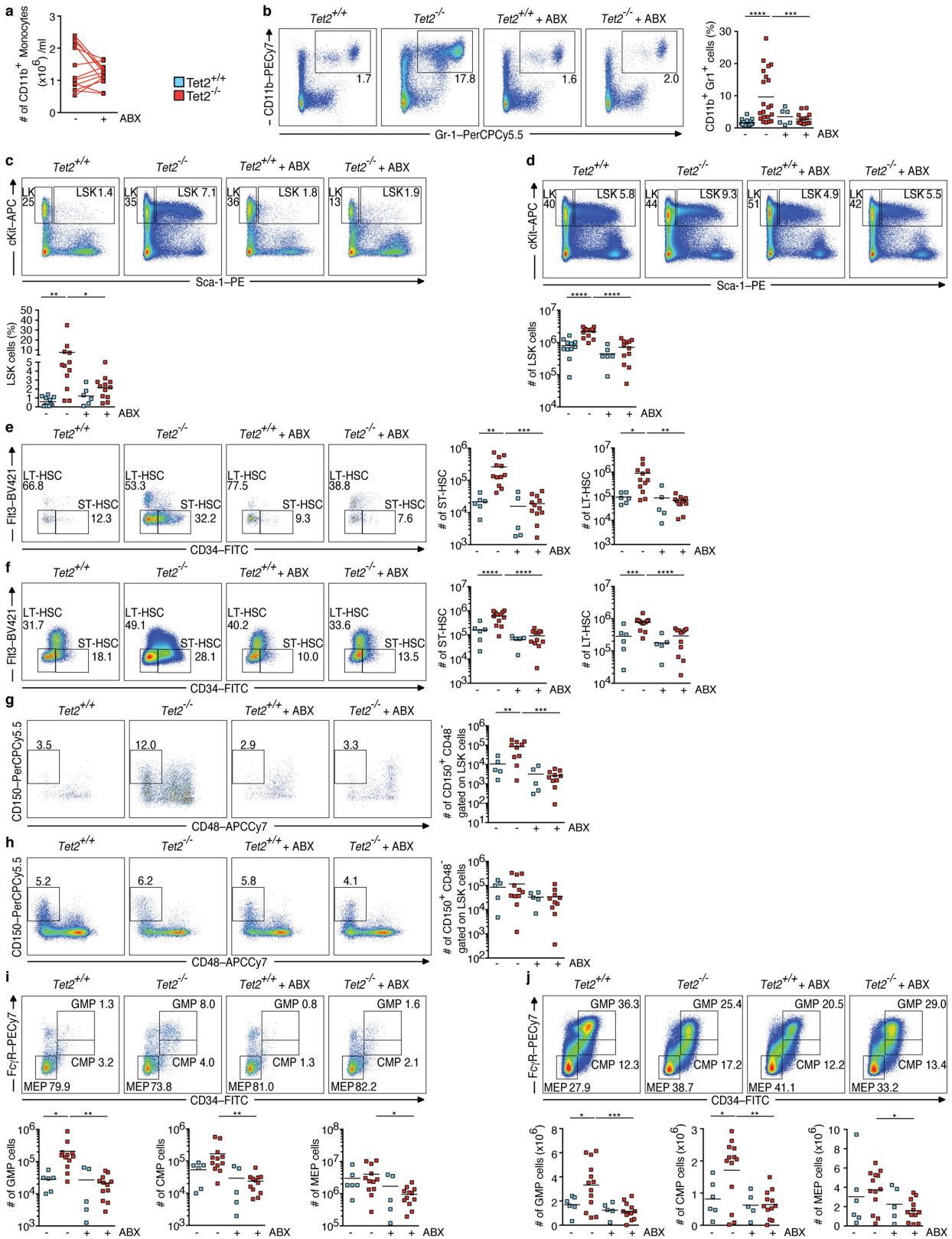
Extended Data Fig. 7 | Microbial community structures are similar between *Tet2*$^{-/-}$ mice and littermate controls. a, **b**, Non-metric multidimensional scaling of samples based on their oligotype profiles (with Bray–Curtis distance). **a**, Microbial community structures within the same body site—that is, jejunum (left, $n = 14$ (*Tet2*$^{-/-}$ (KO)) or 9 mice (*Tet2*$^{+/+}$(WT))), colon (middle, $n = 7$ (KO) or 5 mice (WT)) and faeces (right, $n = 6$ (KO) or 5 mice (WT))—do not differ significantly across genotypes based on envfit test with 999 permutations ($P > 0.05$, two-tailed). **b**, Microbial community structures within the same genotype—WT (top, $n = 5$ (colonic sample (COL) or faecal sample (FEC)) or 9 mice (jejunal sample (JEJ)) and KO (bottom, $n = 6$ (FEC), 7 (COL) or 14 mice (JEJ))—significantly differ across body sites based on the same test ($P < 0.001$). **c**, **d**, Heat-map displays of oligotypes across samples. Clustering dendrograms are computed with Bray–Curtis distance and average linkage algorithm using the oligotype profiles. The intensity of colours indicates the per cent abundance of a given oligotype (rows) in a given sample (columns). **c**, Comparison of genotypes for samples collected from a single body site. **d**, Comparison of genotypes for samples collected across body sites. **e**, Co-housing does not affect the development of PMP in *Tet2*$^{f/f}$*Vav*^{cre} mice. Representative cage configurations of 3 litters from *Tet2*$^{f/f}$*Vav*^{cre} mice with PMP that are over 20 weeks old (red), symptom-free *Tet2*$^{f/f}$*Vav*^{cre} mice (grey), and littermate controls (blue). CON, luminal content; SCR, scraping samples.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 8 | Microbial signals are required for PMP in *Tet2*$^{-/-}$ mice. a, **b**, Germ-free, SPF-housed *Tet2*$^{-/-}$ mice that are over 40 weeks old, and germ-free wild-type controls, analysed for percentage of lymphocytes (**a** left), numbers of lymphocytes (**a** middle left), numbers of CD11b^{+} monocytes (**a** middle right) and numbers of leukocytes (WBC) (**a** right) ($n = 6$ (*Tet2*$^{+/+}$, GF), 7 (*Tet2*$^{-/-}$, GF) or 7 (*Tet2*$^{-/-}$, SPF) mice), and percentage of CD11b^{+}Gr1^{+} myeloid cells (**b** left) and numbers of GMP cells (**b** right), ($n = 7$ (*Tet2*$^{+/+}$, GF), 4 (*Tet2*$^{-/-}$, GF) or 5 (*Tet2*$^{-/-}$, SPF) mice). **c**, Mice treated with antibiotics (ABX) before onset of PMP (see schematic in Fig. 3c. Numbers of CD11b^{+}Gr1^{+} myeloid cells (left; $n = 14$ (*Tet2*$^{+/+}$, no ABX), 11 (*Tet2*$^{-/-}$, no ABX), 7 (*Tet2*$^{+/+}$, with ABX) or 7 (*Tet2*$^{-/-}$, with ABX) mice) and LSK cells (right; $n = 12$ (*Tet2*$^{+/+}$, no ABX), 8 (*Tet2*$^{-/-}$, no ABX), 7 (*Tet2*$^{+/+}$, with ABX) or 7 (*Tet2*$^{-/-}$, with ABX) mice). **d**, *Tet2*$^{-/-}$ mice monitored for the number of CD11b^{+} monocytes (left, $n = 7$ mice) and 16S gene copies in the faeces (right, $n = 6$ mice) before, during and after antibiotics treatment. Mean ± s.e.m., repeated measures one-way ANOVA, Sidak's post hoc test. In **a**–**c**, centre is mean, one-way ANOVA, Sidak's post hoc test, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. Data are representative of at least two independent experiments.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

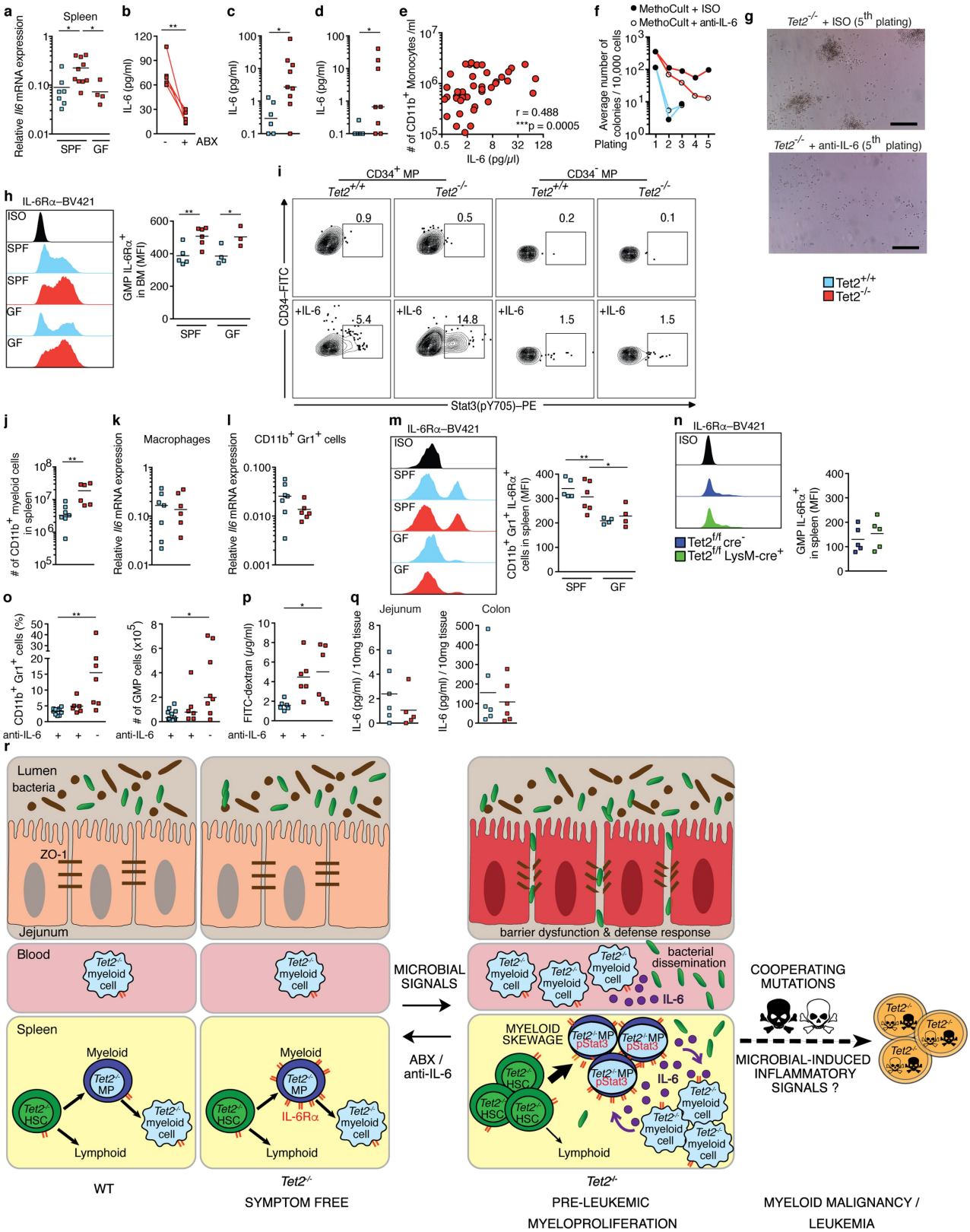


Extended Data Fig. 9 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 9 | Antibiotic treatment reverses PMP in *Tet2*$^{-/-}$ mice. a–j, *Tet2*$^{-/-}$ mice with PMP that are over 20 weeks old, and littermate controls, were treated with and without antibiotics (ABX) for four weeks. **a**, Numbers of peripheral-blood CD11b^{+} monocytes ($n = 13$ mice). Lines connect values obtained from the same mouse sampled before (−) and after (+) ABX treatment. **b**, **c**, Representative dot blots and percentage of CD11b^{+}Gr1^{+} myeloid cells (**b**) ($n = 20$ (*Tet2*$^{+/+}$, no ABX), 21 (*Tet2*$^{-/-}$, no ABX), 6 (*Tet2*$^{+/+}$, with ABX) or 14 (*Tet2*$^{-/-}$, with ABX) mice) and LSK cells in the spleen (**c**) ($n = 11$(*Tet2*$^{+/+}$, no ABX), 12 (*Tet2*$^{-/-}$, no ABX), 6 (*Tet2*$^{+/+}$, with ABX) or 12 (*Tet2*$^{-/-}$, with ABX) mice). **d**, Representative dot blots and numbers of bone marrow-derived LSK cells ($n = 11$(*Tet2*$^{+/+}$, no ABX), 13 (*Tet2*$^{-/-}$, no ABX), 6 (*Tet2*$^{+/+}$, with ABX) or 12 (*Tet2*$^{-/-}$, with ABX) mice). **e**, **f**, Representative dot blots and numbers of splenic (**e**) ($n = 6$ (*Tet2*$^{+/+}$, no ABX), 12 (*Tet2*$^{-/-}$, no ABX), 5 (*Tet2*$^{+/+}$, with ABX) or 12 (*Tet2*$^{-/-}$, with ABX) mice) and bone marrow-derived (**f**) LSK gated CD34^{+}Flt3^{-} (Flt3 is also known as CD135) short-term (ST)-HSCs and CD34^{-}Flt3^{-} long-term (LT)-HSCs ($n = 6$ (*Tet2*$^{+/+}$, no ABX), 13 (*Tet2*$^{-/-}$, no ABX), 5 (*Tet2*$^{+/+}$, with ABX) or 12 (*Tet2*$^{-/-}$, with ABX) mice). **g**, **h**, Representative dot blots and numbers of splenic (**g**) ($n = 5$(*Tet2*$^{+/+}$, no ABX), 10 (*Tet2*$^{-/-}$, no ABX), 5 (*Tet2*$^{+/+}$, with ABX) or 10 (*Tet2*$^{-/-}$, with ABX) mice) and bone marrow-derived (**h**) LSK gated CD150^{+} CD48^{-} cells ($n = 5$ (*Tet2*$^{+/+}$, no ABX), 11 (*Tet2*$^{-/-}$, no ABX), 5 (*Tet2*$^{+/+}$, with ABX) or 10 (*Tet2*$^{-/-}$, with ABX) mice). **i**, **j**, Representative dot blots and numbers of splenic (**i**) ($n = 6$ (*Tet2*$^{+/+}$, no ABX), 12 (*Tet2*$^{-/-}$, no ABX), 5 (*Tet2*$^{+/+}$, with ABX) or 12 (*Tet2*$^{-/-}$, with ABX) mice) and bone marrow-derived (**j**) c-Kit^{+}Sca-1^{-} (LK gated) GMP, common myeloid progenitor (CMP) and megakaryocyte–erythroid progenitor (MEP) cells ($n = 6$ (*Tet2*$^{+/+}$, no ABX), 13 (*Tet2*$^{-/-}$, no ABX), 5 (*Tet2*$^{+/+}$, with ABX) or 12 (*Tet2*$^{-/-}$, with ABX) mice). In **b**–**j**, centre is mean, one-way ANOVA, Sidak's post hoc test. Data are representative of at least three independent experiments. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 10 | Bacteria-induced IL-6 is required for PMP in *Tet2*$^{-/-}$ mice. a, Gene expression of *Il6* in the spleen of SPF-housed *Tet2*$^{-/-}$ mice, littermates and germ-free *Tet2*$^{-/-}$ mice ($n = 7$ (*Tet2*$^{+/+}$, SPF), 10 (*Tet2*$^{-/-}$, SPF) or 4 (*Tet2*$^{-/-}$, GF) mice). **b**, IL-6 cytokine levels in blood plasma of *Tet2*$^{-/-}$ mice with PMP treated with antibiotics (ABX) for four weeks ($n = 5$ mice). Lines connect values obtained from the same mouse sampled before and after antibiotics treatment. Two-tailed paired *t*-test. **c**, IL-6 cytokine levels in blood plasma of DSS-treated symptom-free *Tet2*$^{f/f}$*Vav*^{cre} mice that are over 20 weeks old, and littermate controls (see schematic in Extended Data Fig. 3b ($n = 6$ (*Tet2*$^{+/+}$) or 9 (*Tet2*$^{-/-}$) mice). **d**, IL-6 cytokine levels in blood plasma of *Tet2*$^{f/f}$*Vav*^{cre} mice that are over 20 weeks old, and littermate controls, treated with the TLR1/2 agonist Pam3CSK4 (see schematic in Extended Data Fig. 6b ($n = 6$ (*Tet2*$^{+/+}$) or 8 (*Tet2*$^{-/-}$) mice). **e**, Correlation between IL-6 cytokine levels in blood plasma of *Tet2*$^{-/-}$ mice and numbers of peripheral-blood CD11b^{+} monocytes ($n = 49$ mice). Pearson correlation test. **f**, **g**, In vitro HSC self-renewal colony-forming assay of haematopoietic progenitors of the spleen from *Tet2*$^{-/-}$ mice (red lines) and littermate controls (blue lines) in the presence of anti-IL-6 antibody or isotype control (ISO) after the first replating ($n = 3$ mice). **g**, Representative images of colonies after the 5th replating. Scale bars, 100 μm. **h**, Representative histogram (left) and quantification of (right) mean fluorescence intensity (MFI) of IL-6Rα^{+}c-Kit^{+}Sca-1^{-} (LK gated) CD34^{+}FcγRIII/II^{+} GMPs from the bone marrow, ($n = 5$ (*Tet2*$^{+/+}$, SPF), 6 (*Tet2*$^{-/-}$, SPF), 4 (*Tet2*$^{+/+}$, GF) or 3 (*Tet2*$^{-/-}$, GF) mice). **i**, Representative flow cytometry plot of Stat3 phosphorylation (pY705) response after 30-min stimulation with 10 ng ml^{-1} IL-6 in splenic c-Kit^{+}Sca-1^{-} (LK gated) CD34^{+} and CD34^{-} myeloid progenitors (MP). **j**, Numbers of CD11b^{+} myeloid cells ($n = 7$ (*Tet2*$^{+/+}$) or 6 (*Tet2*$^{-/-}$) mice). **k**, **l**, CD11b^{+}F4/80^{+} macrophages and CD11b^{+}Gr1^{+} myeloid cells of the spleen, from *Tet2*$^{-/-}$ mice and littermate controls, were FACS sorted. Gene expression of *Il6* in macrophages (**k**) and CD11b^{+}Gr1^{+} cells (**l**) ($n = 7$ (*Tet2*$^{+/+}$) or 6 (*Tet2*$^{-/-}$) mice). Centre is median. **m**, Representative histogram (left) and quantification of (right) MFI of IL-6Rα^{+}CD11b^{+}Gr1^{+} myeloid cells in the spleen ($n = 5$ (*Tet2*$^{+/+}$, SPF), 6 (*Tet2*$^{-/-}$, SPF), 4 (*Tet2*$^{+/+}$, GF) or 4 (*Tet2*$^{-/-}$, GF) mice). **n**, Representative histogram (left) and quantification of (right) MFI of IL-6Rα^{+}c-Kit^{+}Sca-1^{-} (LK gated) CD34^{+}FcγRIII/II^{+} GMPs from the spleen of *Tet2*$^{f/f}$*LysM*^{cre} mice and littermate controls ($n = 5$ mice). Centre is mean. **o**, **p**, anti-IL-6 antibody (+) or ISO treatment (−) of *Tet2*$^{-/-}$ mice with PMP that are over 20 weeks old, and littermate controls (see schematic in Fig. 4g. **o**, Percentage of CD11b^{+}Gr1^{+} myeloid cells (left; $n = 10$ (*Tet2*$^{+/+}$, with anti-IL-6), 6 (*Tet2*$^{-/-}$, with anti-IL-6) or 7 (*Tet2*$^{-/-}$, without anti-IL-6) mice) and numbers of GMPs (right; $n = 11$ (*Tet2*$^{+/+}$, with anti-IL-6), 6 (*Tet2*$^{-/-}$, with anti-IL-6) or 8 (*Tet2*$^{-/-}$, without anti-IL-6) mice) in the spleen. **p**, Intestinal permeability was assessed by blood plasma FITC–dextran concentrations ($n = 6$ mice in all cases). **q**, IL-6 in supernatants from intestinal explants: jejunum (left; $n = 6$ (*Tet2*$^{+/+}$) or 5 (*Tet2*$^{-/-}$) mice) and colon (right; $n = 6$ mice in both cases). Centre is mean. **a**, **h**, **m**, **p**, Centre is mean, one-way ANOVA, Sidak's post hoc test. **c**, **d**, **j**, Centre is median, two-tailed Mann–Whitney *U*-test. **o**, Centre is median, Kruskal–Wallis, Dunn's post hoc test. Data are representative of at least three independent experiments; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. **r**, Model showing that extrinsic (microbial-induced inflammatory) and intrinsic (IL-6Rα expression) signals are required for PMP in *Tet2*$^{-/-}$ mice. In detail, small-intestinal barrier dysfunction (reduced ZO-1 and upregulation of defence response genes), which occurs spontaneously or upon intestinal damage, leads to bacterial translocation and to high levels of IL-6. Bacterial translocation can be bypassed when *Tet2*-deficient mice receive systemic microbial signals. Microbial-induced IL-6 is sensed by *Tet2*$^{-/-}$ myeloid progenitor (MP) cells that overexpress IL-6Rα and are highly sensitive to IL-6 (Stat3 (pY705)). Subsequently, MPs expand upon IL-6 signals and preferentially differentiate into mature myeloid cells with IL-6-producing capacities. This cycle results in the development of PMP. Treatment with antibiotics or neutralizing anti-IL-6 antibody can revert PMP, indicating that microbial inflammatory signals are required for PMP in the context of *Tet2* deficiency. However, whether bacteria-induced inflammatory signals also create a permissive environment that induces the acquisition of cooperative oncogenic mutations that lead to the development of leukaemia remains to be determined. The mechanisms through which *Tet2* deficiency in haematopoietic cells leads to a microbiota-dependent impairment of gut barrier function remains to be addressed.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0128-9

KLHL22 activates amino-acid-dependent mTORC1 signalling to promote tumorigenesis and ageing

Jie Chen[1,2], Yuhui Ou[1], Yanyan Yang[1], Wen Li[3], Ye Xu[4], Yuntao Xie[4] & Ying Liu[1]*

The mechanistic target of rapamycin complex 1 (mTORC1) is a master regulator of cell growth that responds to a diverse set of environmental cues, including amino acids[1,2]. Deregulation of mTORC1 has been linked with metabolic diseases, cancer and ageing[2–4]. In response to amino acids, mTORC1 is recruited by the Rag GTPases to the lysosome, its site of activation[5,6]. The GATOR1 complex, consisting of DEPDC5, NPRL3 and NPRL2, displays GAP activity to inactivate Rag GTPases under amino-acid-deficient conditions[7]. However, it is unclear how the inhibitory function of GATOR1 is released upon amino acid stimulation. Here we find that in response to amino acids, the CUL3–KLHL22 E3 ubiquitin ligase promotes K48-linked polyubiquitination and degradation of DEPDC5, an essential subunit of GATOR1. KLHL22 plays a conserved role to mediate the activation of mTORC1 and downstream events in mammals and nematodes. Depletion of MEL-26, the *Caenorhabditis elegans* orthologue of KLHL22, extends worm lifespan. Moreover, KLHL22 levels are elevated in tumours of breast cancer patients, whereas DEPDC5 levels are correspondingly reduced. Depletion of KLHL22 in breast cancer cells suppresses tumour growth in nude mice. Therefore, pharmacological interventions targeting KLHL22 may have therapeutic potential for the treatment of breast cancer and age-related diseases.

To understand how inhibition by GATOR1 is released during mTORC1 activation, we monitored the stability of GATOR1 subunits in response to amino acid availability. Notably, the protein levels of DEPDC5, but not NPRL3 or NPRL2, were regulated in an amino-acid-sensitive manner (Fig. 1a and Extended Data Fig. 1a). By contrast, levels of DEPDC5 transcripts remained unchanged in the presence or absence of amino acids (Extended Data Fig. 1b). The stability of DEPDC5 protein was extremely low under the basal culturing condition (Extended Data Fig. 1c, d). Accumulation of DEPDC5 was detected only when we treated cells with MG132, the 26S proteasome inhibitor (Extended Data Fig. 1e), suggesting that DEPDC5 undergoes proteasome-mediated degradation in response to amino acids.

Covalent conjugation of ubiquitin is a key step in proteasome-mediated degradation of target proteins. Indeed, the ubiquitination of DEPDC5 was observed in the presence of amino acids (Fig. 1b and Extended Data Fig. 1f). DEPDC5 could be labelled only with wild-type or K48 ubiquitin (ubiquitin mutant that contains only one lysine), but not with K63 ubiquitin (Fig. 1c), indicating that amino acids promote K48-linked ubiquitination of DEPDC5.

We next sought to identify the E3 ubiquitin ligase that targets DEPDC5. Because mTORC1 signalling is often deregulated in human cancers[8,9], and the GATOR1 complex has been reported to function on the lysosomal surface[10], we screened a panel of E3 ligases that have been reported to affect mTORC1 activity or tumorigenesis, together with E3 ligases localized to the lysosomes. We found that ectopic expression of KLHL22, mutations of which have been linked with breast cancer[11], promoted the degradation of endogenous DEPDC5 (Extended Data Fig. 2a). KLHL22 is a BTB (Bric-á-brac–Tramtrack–Broad) adaptor protein, usually forming a functional cullin–RING E3 ubiquitin ligase complex with the scaffold protein CUL3 and the ring-finger protein RBX1[12,13]. Notably, CUL3 and RBX1 were identified by mass spectrometry during our search for DEPDC5-interacting proteins (Extended Data Fig. 2b). The substrate specificity of cullin–RING E3 ubiquitin ligase is determined through a BTB adaptor protein within the complex[14]. Indeed, overexpression of CUL3–RBX1–KLHL22, but not CUL3–RBX1–KLHL19, promoted the ubiquitination and degradation of DEPDC5 (Fig. 1d, e). Deletion of the 6-Kelch (6K) repeats, the substrate recognition motif of KLHL22[15], or knockdown of *KLHL22* by small interfering RNA (siRNA), blocked DEPDC5 ubiquitination and degradation (Fig. 1d–g). Furthermore, recombinant CUL3–RBX1–KLHL22 proteins were sufficient to promote DEPDC5 ubiquitination in vitro (Extended Data Fig. 2c, d).

Several lysine residues were predicted by UbPred[16], or have previously been identified by mass spectrometry as possible ubiquitination sites of DEPDC5[17] (Extended Data Fig. 3a). K48-linked polyubiquitination and degradation of DEPDC5, and subsequent mTORC1 activation (indicated by S6K1 phosphorylation), were blocked if all five lysine residues were simultaneously mutated to arginine (5KR) (Extended Data Fig. 3b, c). However, none of the single mutations blocked DEPDC5 ubiquitination and degradation (Extended Data Fig. 3d, e). Therefore, CUL3–KLHL22 E3 ligase catalyses K48-linked ubiquitination on multiple sites of DEPDC5.

We next investigated how KLHL22 regulates DEPDC5 in response to amino acid availability. KLHL22 interacted specifically with DEPDC5 in an amino-acid-sensitive manner (Fig. 1h and Extended Data Fig. 4a–c). The direct interaction between DEPDC5 and KLHL22 was further demonstrated using recombinant proteins in an in vitro binding assay (Extended Data Fig. 4d). However, the DEPDC5(5KR) mutant was not able to interact with KLHL22 (Extended Data Fig. 4d), possibly owing to the aberrant folding of DEPDC5 caused by the mutations. Therefore, the lysine residues responsible for DEPDC5 ubiquitylation by KLHL22 warrant further investigation. Through truncation mapping tests, the DEP domain was identified as the degron of DEPDC5 (Extended Data Fig. 4e, f), which is responsible for KLHL22-mediated degradation (Extended Data Fig. 4g). Mutation of each serine, threonine, or tyrosine within the degron did not prevent its interaction with KLHL22 (Extended Data Fig. 4h), suggesting that it might not be a phospho-degron.

To understand how amino acids modulate KLHL22 activity, we monitored the localization of KLHL22 in response to amino acids. Notably, amino acids mediated nuclear–cytosolic shuttling of KLHL22 (Extended Data Fig. 5a, b). Several phosphorylation sites have been reported in KLHL22[11] (Extended Data Fig. 5c). We mutated each site to alanine, and found that S18 was required for the nuclear accumulation of KLHL22 in amino-acid-deprived conditions (Extended Data Fig. 5d). Through mass spectrometry analysis, we found that KLHL22 associated with 14-3-3 proteins during amino acid starvation (Extended Data Fig. 5e). The S18A mutation disrupted the interaction between

[1]State Key Laboratory of Membrane Biology, Institute of Molecular Medicine, Peking-Tsinghua Center for Life Sciences, Beijing Key Laboratory of Cardiometabolic Molecular Medicine, Peking University, Beijing, China. [2]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China. [3]College of Biological Science and Technology, Beijing Forestry University, Beijing, China. [4]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Breast Center, Peking University Cancer Hospital & Institute, Beijing, China.
*e-mail: ying.liu@pku.edu.cn

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

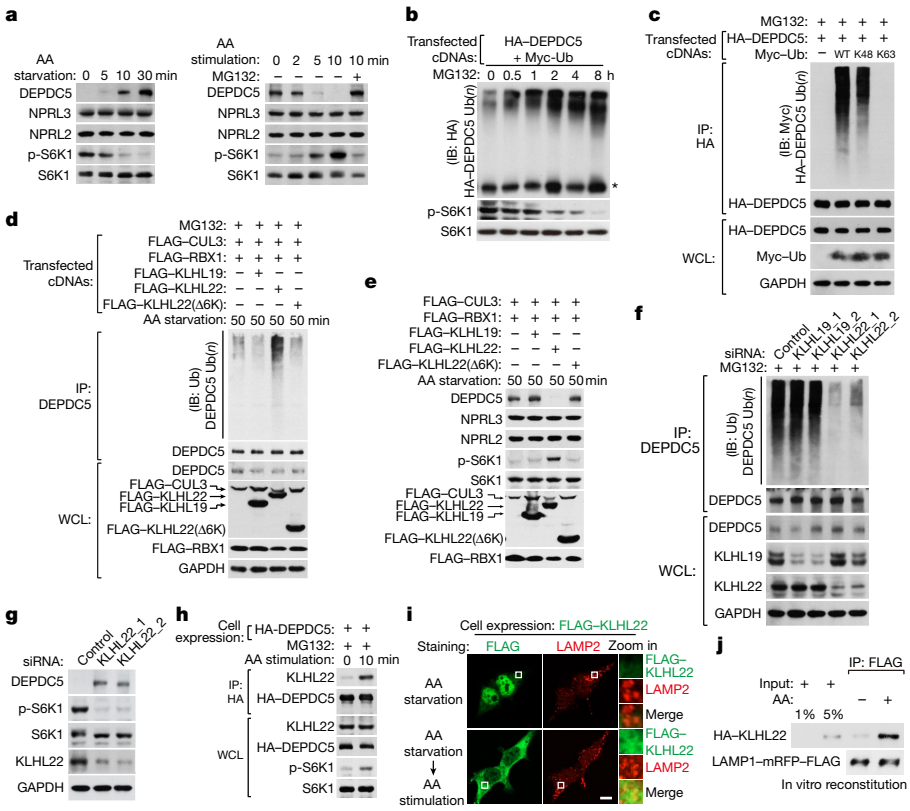


Fig. 1 | CUL3–KLHL22 mediates K48-linked polyubiquitination and degradation of DEPDC5. a, DEPDC5 is regulated in an amino-acid-sensitive manner. AA, amino acid. **b**, Accumulation of ubiquitinated DEPDC5 upon MG132 treatment. The asterisk represents HA–DEPDC5. IB, immunoblot; Ub(*n*), polyubiquitinated. **c**, DEPDC5 is modified by K48-linked polyubiquitination. IP, immunoprecipitation; WCL, whole-cell lysate. **d**, **e**, CUL3–KLHL22 promotes the ubiquitination (**d**) and degradation (**e**) of DEPDC5. **f**, **g**, Knockdown of KLHL22 suppresses the ubiquitination (**f**) and degradation (**g**) of DEPDC5. **h**, KLHL22 interacts with DEPDC5 in an amino-acid-dependent manner. **i**, KLHL22 partially locates on lysosomes under amino-acid-sufficient conditions. LAMP2, lysosome. Scale bar, 10 μm. **j**, Recombinant KLHL22 accumulates on purified lysosomes in response to amino acids. For gel source data, see Supplementary Fig. 1.

KLHL22 and 14-3-3, leading to the constitutive activation of mTORC1 (Extended Data Fig. 5f, g). Therefore, during amino acid starvation, KLHL22 was trapped by 14-3-3 proteins in the nucleus. Upon amino acid stimulation, KLHL22 was released and translocated to the cytosol. Once in the cytosol, KLHL22 localized at least in part to the lysosome (Fig. 1i), where GATOR1 resides[10,18]. Consistently, KLHL22 accumulated on purified lysosomes in the in vitro reconstitution system when amino acids were supplemented (Fig. 1j and Extended Data Fig. 5h, i).

Recruitment of mTORC1 to the surface of lysosomes is a key step for amino-acid-induced mTORC1 activation[5]. In KLHL22 knockout cells (sgKLHL22), mTORC1 could not accumulate on lysosomes upon amino acid stimulation (Fig. 2a and Extended Data Fig. 6a, b). In addition, in KLHL22 knockout cells, amino acid replenishment was not able to reduce the DEPDC5 level or activate mTORC1 (Fig. 2b). Conversely, ectopic expression of CUL3–KLHL22 induced S6K1 phosphorylation (Fig. 2c). Overexpression or depletion of KLHL22 did not perturb mTORC1 signalling in DEPDC5-deficient cells (Extended Data Fig. 6c, d), further suggesting that KLHL22 regulates the mTORC1 pathway through modulation of DEPDC5.

Expression of dominant-negative Rag GTPases (Rag^{DN})[5,6] completely blocked KLHL22-mediated mTORC1 activation (Fig. 2c), indicating that KLHL22 acts upstream of Rag GTPases to mediate mTORC1 signalling. Consistently, KLHL22 displayed nuclear–cytosolic shuttling and activated mTORC1 in response to leucine but not glutamine, an amino acid that activates mTORC1 in a Rag-independent fashion[19] (Fig. 2d, e). mTORC1 negatively regulates catabolic pathways such as autophagy[20]. Deletion of KLHL22 activated autophagy, as determined by the increased LC3BII/I ratio (Fig. 2f and Extended Data Fig. 6e) and the induction of LC3B puncta (Fig. 2g and Extended Data Fig. 6f). KLHL22 deficiency also reduced cell size (Fig. 2h), a phenotypic characteristic of mTORC1 inactivation[21]. Thus, KLHL22 is essential for the activation of mTORC1and downstream events.

To investigate whether KLHL22 plays an evolutionarily conserved role in TORC1 signalling, we knocked out KLHL22 in mouse embryonic fibroblast (MEF) cells, and showed that mTORC1 failed to accumulate on lysosomes and phosphorylate S6K1 upon amino acid stimulation (Fig. 3a, b and Extended Data Fig. 7a). Multiple genes are predicted as potential *KLHL22* orthologues in *C. elegans* (Extended Data Fig. 7b). We screened these genes with the use of a reporter strain expressing GFP-tagged HLH-30, an orthologue of human transcription factor EB (TFEB) that translocalizes to the nucleus upon TORC1 inhibition[22,23]. Knockdown of *mel-26* or *tag-53*, but not any other gene, induced the nuclear accumulation of HLH-30 (Fig. 3c and Extended Data Fig. 7c). RNAi knockdown of *T08A11.1*, which encodes the orthologue of human DEPDC5, impaired HLH-30 nuclear localization in *mel-26*-deficient animals, but not in *tag-53*-deficient worms (Fig. 3d and Extended Data Fig. 7c). Downregulation of *T08A11.1* also restored the developmental delay of *mel-26*-deficient worms (Extended Data Fig. 7d). More importantly, expression of codon-optimized human KLHL22 suppressed the nuclear accumulation of HLH-30 in *mel-26* mutant animals (Fig. 3e), suggesting that MEL-26 functions as the *C. elegans* orthologue of KLHL22 to regulate T08A11.1 (*Ce*.DEPDC5).

Inhibition of TORC1 has been linked with lifespan extension in several species[24–27]. To determine the physiological significance of KLHL22, we compared the lifespans of wild-type worms with those of *mel-26* mutants. Deletion of *mel-26* significantly extended worm lifespan (Fig. 3f). The prolonged lifespan of *mel-26* mutants was suppressed by RNAi knockdown of *T08A11.1* (Fig. 3g).

The mTORC1 pathway is frequently deregulated in many cancer types[8,9]. Deletions in GATOR1 subunits NPRL2 and DEPDC5 have

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

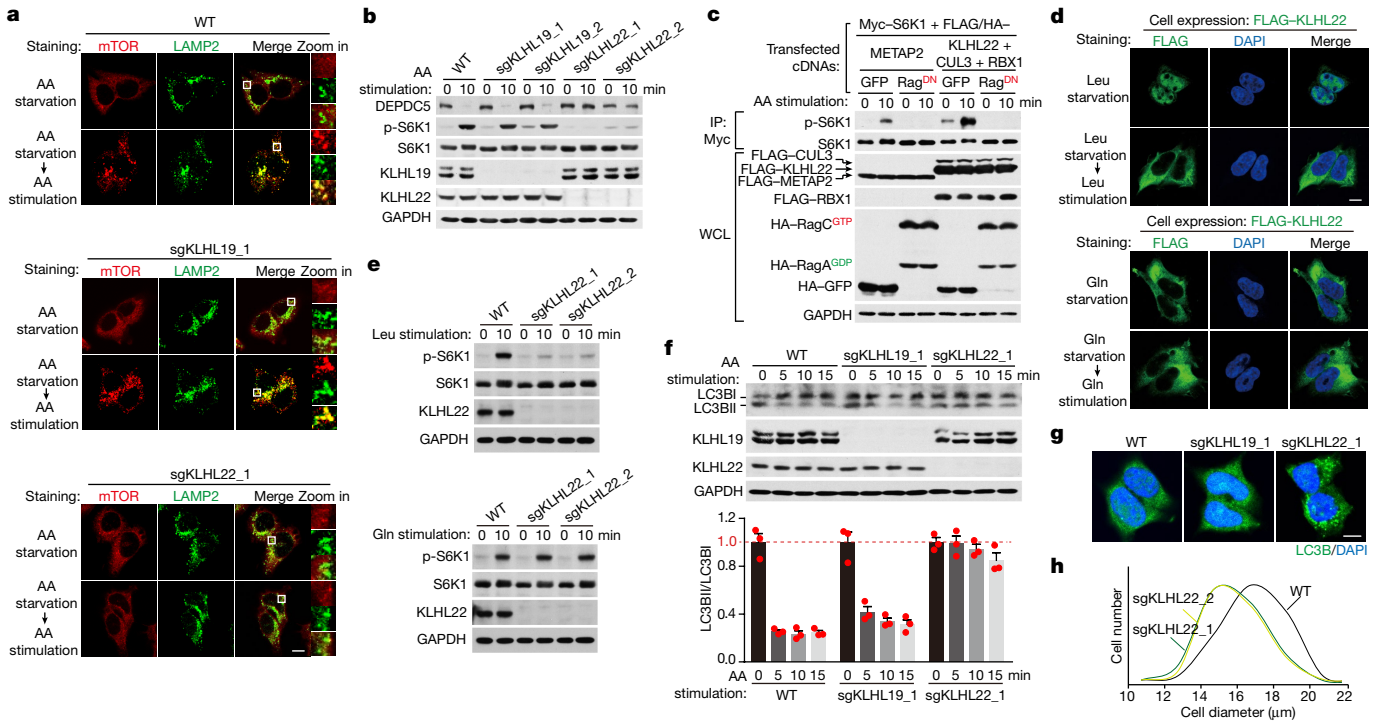


Fig. 2 | KLHL22 is required for mTORC1 activation, autophagy and cell growth. a, KLHL22 depletion prevents amino-acid-stimulated lysosomal accumulation of mTORC1. Scale bar, 10 μm. **b**, Depletion of KLHL22 suppresses amino-acid-dependent degradation of DEPDC5 and phosphorylation of S6K1. WT, wild-type. **c**, KLHL22 plays a role upstream of Rag GTPases. Rag^{DN}: $RagA^{GDP}$–$RagC^{GTP}$, dominant-negative form of Rag GTPases. **d**, FLAG–KLHL22 translocates to the cytosol upon stimulation with leucine but not glutamine. DAPI, nuclei. Scale bar, 10 μm. **e**, Deprivation of KLHL22 prevents leucine-induced but not glutamine-induced S6K1 phosphorylation. **f**, KLHL22 is required for amino-acid-induced suppression of autophagy. $n = 3$ biologically independent experiments. **g**, Depletion of KLHL22 induces formation of LC3B puncta. DAPI, nuclei. Scale bar, 10 μm. **h**, Depletion of KLHL22 reduces cell size. For gel source data, see Supplementary Fig. 1.

been reported in lung and breast cancers, and two cases of glioblastoma[28–30]. Overexpression of DEPDC5 reduced cell size (Extended Data Fig. 8a), whereas co-expression of KLHL22 restored it (Extended Data Fig. 8b). Furthermore, a stable HEK293T cell line with a high level of KLHL22 showed greatly altered growth morphology and increased anchorage-independent growth (Fig. 4a, b and Extended Data Fig. 8c). High-level expression of KLHL22 conferred increased sensitivity to rapamycin, an mTORC1 inhibitor (Fig. 4b).

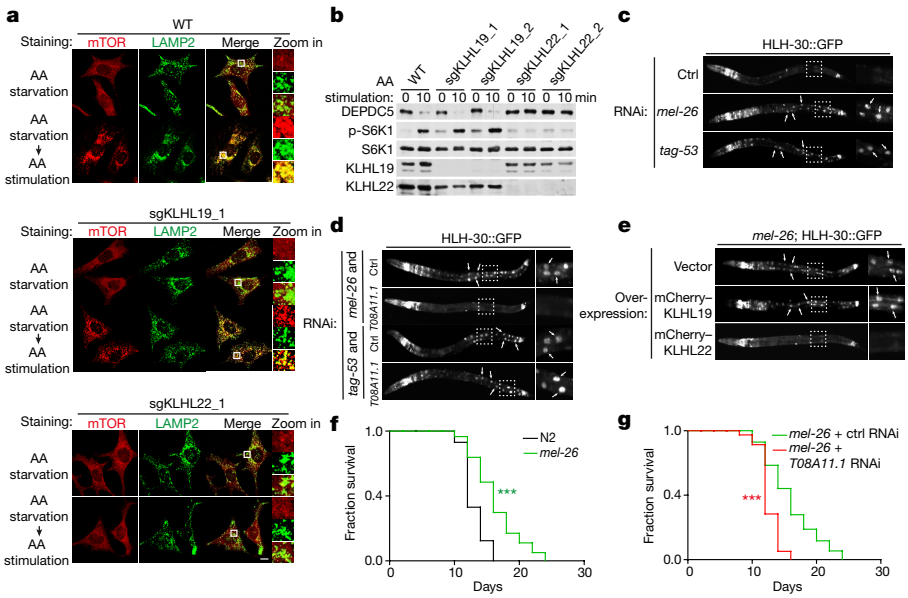


Fig. 3 | KLHL22 plays an evolutionarily conserved role in the regulation of DEPDC5. a, KLHL22 is required for amino-acid-induced lysosomal localization of mTOR in MEF cells. Scale bar, 10 μm. **b**, Deficiency of KLHL22 suppresses amino-acid-induced degradation of DEPDC5 and phosphorylation of S6K1 in MEF cells. For gel source data, see Supplementary Fig. 1. **c**, RNA interference (RNAi) targeting *mel-26* or *tag-53* induces nuclear accumulation of HLH-30::GFP in *C. elegans* (arrows). Dashed boxes are expanded on right. **d**, RNAi targeting *T08A11.1* suppresses nuclear accumulation of HLH-30 in *mel-26*-deficient, but not *tag-53*-deficient worms. **e**, Human KLHL22 prevents the nuclear accumulation of HLH-30 in *mel-26*-deficient worms. **f**, **g**, *mel-26* mutant worms have prolonged lifespan (**f**) and RNAi of *T08A11.1* suppresses lifespan extension in *mel-26* mutant worms (**g**). $n = 2$ biologically independent experiments. $^{***}P < 0.001$; log rank test of one replicate is shown.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

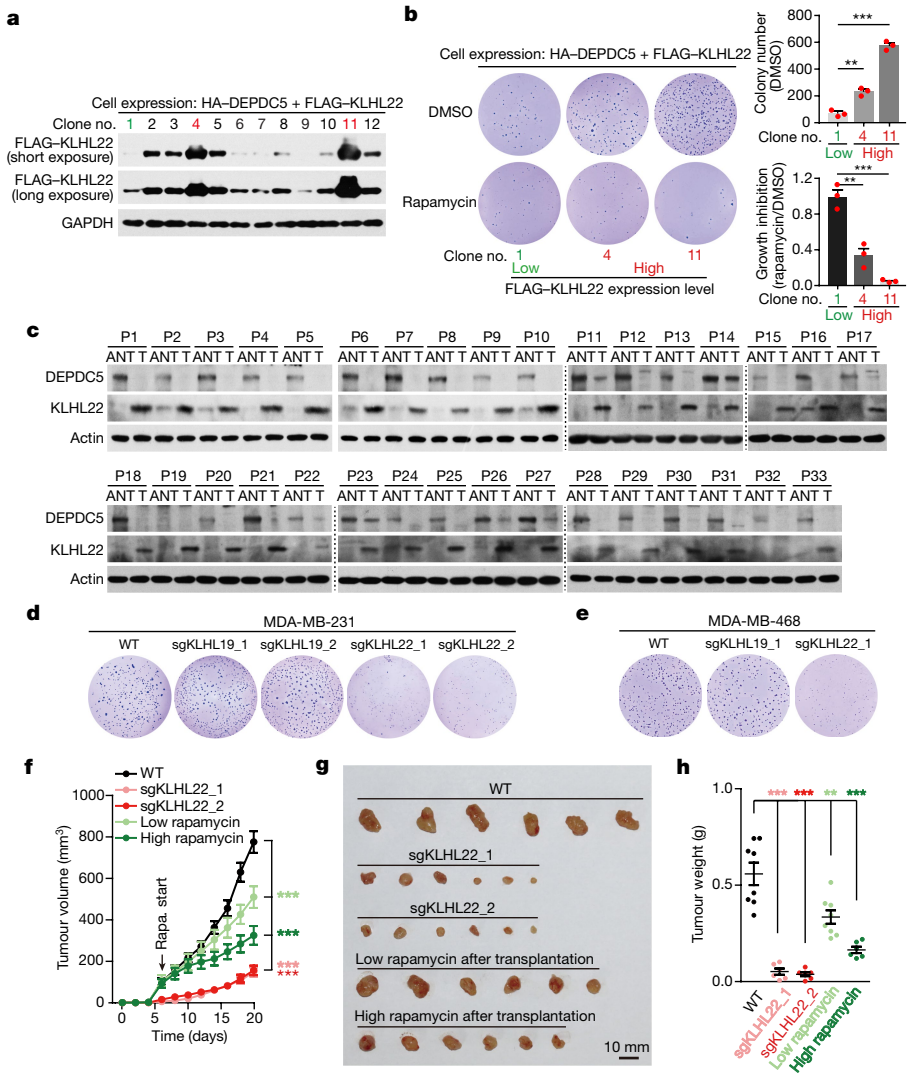


Fig. 4 | KLHL22 acts as a potential oncogene in breast cancer. a, Protein levels of FLAG–KLHL22 in cells stably expressing HA–DEPDC5 and FLAG–KLHL22. Green or red denotes low or high protein level of KLHL22. **b**, High levels of KLHL22 promote anchorage-independent growth and confer increased rapamycin sensitivity. $n = 3$ biological replicates. $^{**}P < 0.01$; $^{***}P < 0.001$, two-sided Student's t-test. **c**, Upregulation of KLHL22 and downregulation of DEPDC5 are detected in tumours of breast cancer patients. T, tumour; ANT, adjacent normal tissues; P1, pair 1. **d**, **e**, Depletion of KLHL22 inhibits anchorage-independent growth of MDA-MB-231 (**d**) and MDA-MB-468 (**e**) human breast cancer cells. **f**–**h**, Deletion of KLHL22 or treatment with rapamycin suppresses tumour growth of MDA-MB-231 cells. Tumour volume, mean ± s.e.m. ($n = 9$ mice for sgKLHL22_2, $n = 10$ mice for remaining groups), $^{***}P < 0.001$, two-sided ANOVA (**f**); tumour images (**g**); and tumour weights 20 d after transplantation, mean ± s.e.m. ($n = 8$ mice for WT and WT injected with low dose of rapamycin, $n = 6$ mice for remaining groups), $^{***}P < 0.001$, two-sided Student's t-test (**h**). For gel source data, see Supplementary Fig. 1.

The 'Oncomine' database revealed several cancer types with elevated KLHL22 transcript levels (Extended Data Fig. 8d). Protein levels of KLHL22 were elevated in multiple breast cancer cell lines, such as MDA-MB-231, MDA-MB-468 and BT549, compared with the corresponding control cell line MCF10A (Extended Data Fig. 8e). More strikingly, the level of KLHL22 protein was markedly increased in breast tumours of patients, in comparison with that of the adjacent normal tissues (Fig. 4c). The elevation of KLHL22 correlated strongly with a reduction in DEPDC5 protein level in each tumour sample (Fig. 4c). KLHL22 mRNA level was also elevated in tumour samples (Extended Data Fig. 8f), suggesting that KLHL22 might be regulated at the transcriptional level in breast cancer.

Breast cancer cells (for example, MDA-MB-231) displayed similar regulation of DEPDC5 in response to amino acids (Extended Data Fig. 8g–i). Thus, to test whether KLHL22 promotes tumour growth, we knocked out KLHL22 in MDA-MB-231 and MDA-MB-468, two breast cancer cell lines with elevated KLHL22 levels (Extended Data Figs 9a–c). Depletion of KLHL22 suppressed amino acid-triggered degradation of DEPDC5 and phosphorylation of S6K1, and inhibited anchorage-independent growth of both cell lines (Fig. 4d, e and Extended Data Fig. 9d, e). Finally, we subcutaneously transplanted wild-type MDA-MB-231 and MDA-MB-468 cells, or KLHL22 knockout cells, into immunodeficient nude mice. Depletion of KLHL22 significantly prevented tumour growth (Fig. 4f–h, Extended Data Fig. 9f–h). Moreover, treatment with rapamycin suppressed tumour growth (Fig. 4f–h, Extended Data Fig. 9i–k). Collectively, these data demonstrate that KLHL22 functions as a potential oncogene to active mTORC1 signalling and promote tumour growth in breast cancer.

In response to amino acids, the inhibitory function of GATOR1 must be released in order to activate mTORC1. The identification of CUL3–KLHL22 E3 ubiquitin ligase provides a novel mechanism for amino-acid-induced GATOR1 inactivation. KLHL22 promotes ageing in *C. elegans*, and tumour growth in mice, adding to the physiological significance of this regulation (Extended Data Fig. 10). Strikingly, tumour samples from breast cancer patients all have elevated KLHL22 protein levels, which is correlated with the reduction in DEPDC5

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

levels. Small molecules that inhibit KLHL22 activity on DEPDC5 are candidates for development as drugs for the treatment of breast cancer and age-related diseases.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0128-9.

Received: 2 November 2017; Accepted: 12 April 2018;
Published online: 16 May 2018

1. Dibble, C. C. & Manning, B. D. Signal integration by mTORC1 coordinates nutrient input with biosynthetic output. *Nat. Cell Biol.* **15**, 555–564 (2013).
2. Laplante, M. & Sabatini, D. M. mTOR signaling in growth control and disease. *Cell* **149**, 274–293 (2012).
3. Yuan, H. X., Xiong, Y. & Guan, K. L. Nutrient sensing, metabolism, and cell growth control. *Mol. Cell* **49**, 379–387 (2013).
4. Zoncu, R., Efeyan, A. & Sabatini, D. M. mTOR: from growth signal integration to cancer, diabetes and ageing. *Nat. Rev. Mol. Cell Biol.* **12**, 21–35 (2011).
5. Kim, E., Goraksha-Hicks, P., Li, L., Neufeld, T. P. & Guan, K. L. Regulation of TORC1 by Rag GTPases in nutrient response. *Nat. Cell Biol.* **10**, 935–945 (2008).
6. Sancak, Y. et al. The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science* **320**, 1496–1501 (2008).
7. Bar-Peled, L. et al. A tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. *Science* **340**, 1100–1106 (2013).
8. Sabatini, D. M. mTOR and cancer: insights into a complex relationship. *Nat. Rev. Cancer* **6**, 729–734 (2006).
9. Guertin, D. A. & Sabatini, D. M. Defining the role of mTOR in cancer. *Cancer Cell* **12**, 9–22 (2007).
10. Wolfson, R. L. et al. KICSTOR recruits GATOR1 to the lysosome and is necessary for nutrients to regulate mTORC1. *Nature* **543**, 438–442 (2017).
11. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).
12. Petroski, M. D. & Deshaies, R. J. Function and regulation of cullin-RING ubiquitin ligases. *Nat. Rev. Mol. Cell Biol.* **6**, 9–20 (2005).
13. Dhanoa, B. S., Cogliati, T., Satish, A. G., Bruford, E. A. & Friedman, J. S. Update on the Kelch-like (KLHL) gene family. *Hum. Genomics* **7**, 13 (2013).
14. Pintard, L., Willems, A. & Peter, M. Cullin-based ubiquitin ligases: Cul3-BTB complexes join the family. *EMBO J.* **23**, 1681–1687 (2004).
15. Beck, J. et al. Ubiquitylation-dependent localization of PLK1 in mitosis. *Nat. Cell Biol.* **15**, 430–439 (2013).
16. Radivojac, P. et al. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* **78**, 365–380 (2010).
17. Wagner, S. A. et al. A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol. Cell. Proteomics* **10**, M111–013284 (2011).
18. Peng, M., Yin, N. & Li, M. O. SZT2 dictates GATOR control of mTORC1 signalling. *Nature* **543**, 433–437 (2017).
19. Jewell, J. L. et al. Metabolism. Differential regulation of mTORC1 by leucine and glutamine. *Science* **347**, 194–198 (2015).
20. Hosokawa, N. et al. Nutrient-dependent mTORC1 association with the ULK1-Atg13-FIP200 complex required for autophagy. *Mol. Biol. Cell* **20**, 1981–1991 (2009).
21. Fingar, D. C., Salama, S., Tsou, C., Harlow, E. & Blenis, J. Mammalian cell size is controlled by mTOR and its downstream targets S6K1 and 4EBP1/eIF4E. *Genes Dev.* **16**, 1472–1487 (2002).
22. Lapierre, L. R. et al. The TFEB orthologue HLH-30 regulates autophagy and modulates longevity in *Caenorhabditis elegans*. *Nat. Commun.* **4**, 2267 (2013).
23. Settembre, C. et al. A lysosome-to-nucleus signalling mechanism senses and regulates the lysosome via mTOR and TFEB. *EMBO J.* **31**, 1095–1108 (2012).
24. Kaeberlein, M. et al. Regulation of yeast replicative life span by TOR and Sch9 in response to nutrients. *Science* **310**, 1193–1196 (2005).
25. Vellai, T. et al. Genetics: influence of TOR kinase on lifespan in *C. elegans*. *Nature* **426**, 620 (2003).
26. Kapahi, P. et al. Regulation of lifespan in *Drosophila* by modulation of genes in the TOR signaling pathway. *Curr. Biol.* **14**, 885–890 (2004).
27. Harrison, D. E. et al. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* **460**, 392–395 (2009).
28. Lerman, M. I. & Minna, J. D. The 630-kb lung cancer homozygous deletion region on human chromosome 3p21.3: identification and evaluation of the resident candidate tumor suppressor genes. The International Lung Cancer Chromosome 3p21.3 Tumor Suppressor Gene Consortium. *Cancer Res.* **60**, 6116–6133 (2000).
29. Li, J. et al. Functional characterization of the candidate tumor suppressor gene NPRL2/G21 located in 3p21.3C. *Cancer Res.* **64**, 6438–6443 (2004).
30. Seng, T. J. et al. Complex chromosome 22 rearrangements in astrocytic tumors identified using microsatellite and chromosome 22 tile path array analysis. *Genes Chromosom. Cancer* **43**, 181–193 (2005).

Acknowledgements We thank S. Perrett and S. Song for proofreading our manuscript. We thank S.-C. Lin, G. Ouyang, Q. Wu and Y. Wang for providing us with reagents. Several *C. elegans* strains were provided by CGC, which is supported by the NIH-Officer of Research Infrastructure Programs. This work is supported by the Ministry of Science and Technology of China (National Key Research and Development Program of China grant 2017YFA0504000 and 973 grant 2013CB910104), the National Natural Science Foundation of China (31422033), Peking-Tsinghua Center for Life Sciences, and an HHMI International Research Scholar Award (HHMI#55008739) to Y.L.

Reviewer information *Nature* thanks F. Basserman and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.C., assisted by Y.O., performed most of the experiments and analysed the data. Y.Y. generated the *mel-26*; HLH-30::GFP strain. W.L. generated DEPDC5 and KLHL22 point mutation constructs. Y.Xu. and Y.Xi. provided breast cancer samples. J.C. and Y.L. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0128-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0128-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Y.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Antibodies. Anti-HA (Santa Cruz 7392), anti-FLAG (Sigma F7425), anti-LAMP2 (IF in HEK293T, Abcam 25631), anti-LAMP2 (IF in MEF, Abcam 13524), anti-mTOR (CST 2983), anti-LC3B (Sigma L7543), anti-Myc (ImmunoWay YM3203), anti-Ub (Santa Cruz 8017), anti-KLHL22 (Proteintech 16214-1-AP), anti-KLHL19 (Proteintech10503-2-AP), anti-NPRL2 (Sigma SAB1305758), anti-NPRL3 (Sigma HPA011741), anti-DEPDC5 (Sigma SAB1302644), anti-p-S6K1 (CST 9205), anti-S6K1 (CST 9202), anti-PDI (Santa Cruz 20132), anti-Prohibitin (Santa Cruz 28259), anti-EEA1 (CST 3288), anti-GAPDH (Abcam 128915), anti-Lamin B1 (Bioworld MB2029), anti-Histone H3 (CST 3638), anti-Actin (Sigma A5441), and anti-14-3-3 (Santa Cruz 23957) were used.

Cell lines and cell culture. MCF10A, MCF7, SKBR3, BT549, PWR1E, 22RV1, LNCAP and HSF cells were provided by G. Ouyang (Xiamen University); A375 and MV3 cells were provided by Q. Wu (Xiamen University); MDA-MD-231, MDA-MB-468, PC3 and MEF cells were gifts from S.-C. Lin (Xiamen University). HEK293T cells were purchased from ATCC. None of the cell lines in this study appears in the misidentified cell line list kept by the ICLAC. Cell lines served in this study were not authenticated immediately before use in our laboratory. All cell lines were validated to be free of mycoplasma contamination.

All cell lines were maintained at 37 °C and 5% CO_2, with the exception of MDA-MB-231 and MDA-MB-468 cells, which were cultured without CO_2. MCF10A cells were maintained in MEBM medium (Lonza) supplemented with its additives (without using GA-1000) and 100 ng/ml cholera toxin (Sigma). MCF7 cells were maintained in MEM medium (Gibco) supplemented with 10% inactivated fetal bovine serum (FBS), penicillin and streptomycin (P/S). PC3, 22RV1, LNCAP and BT549 cells were kept in RPMI1640 medium (Gibco) supplemented with 10% FBS and penicillin/streptomycin (P/S). PWR1E, A375, HSF, MV3, MEF and HEK293T cells and their derivatives were maintained in DMEM high glucose medium (Hyclone) supplemented with 10% FBS and P/S. SKBR3 cells were cultured in McCoy's 5a medium (Gibco) supplemented with 10% FBS and P/S. MDA-MB-231 and MDA-MB-468 cells and their derivatives were kept in L15 medium (Gibco) supplemented with 10% FBS and P/S. Except where indicated, data were generated using HEK293T cells.

For amino acid (AA), leucine or glutamine starvation, cells were incubated in handmade DMEM base medium lacking AAs, leucine or glutamine for the indicated times. For AA, leucine or glutamine stimulation, cells were starved for 50 min and then stimulated by directly adding AAs, leucine or glutamine high concentration stock solution for the indicated times. For starvation and re-stimulation experiments in Fig. 1a and Extended Data Fig. 1f, 5a, dialysed FBS (Gibco, 26400044) was supplemented. For starvation and stimulation experiments in the remaining figures, insulin (Sigma, I9278) was supplemented at 200 ng/ml. High concentration AA solution is a combination of commercial 50 × glutamine-free AA mixture (Gibco, 11130051) and 50 × glutamine (Gibco, 25030081). High concentration leucine and glutamine solutions were handmade.

For drug treatments, MG132 dissolved in DMSO (Santa Cruz 201270) was pre-diluted with medium to 10 μg/ml. Medium containing MG132 was then used to replace the original medium and cells were cultured in the presence of MG132 for 1 h. When supplemented along with amino acids, MG132 was continuously present during starvation, or starvation and re-stimulation periods. Chloroquine (CQ) (Sigma C6628) or cycloheximide (CHX) (Sigma, C7698) dissolved in water was added directly into the culturing medium at a final concentration of 50 μM or 50 μg/ml, respectively.

Nematode strains and culture conditions. *C. elegans* strains (N2; EU1007: *mel-26*; MAH240: HLH-30::GFP) were obtained from the *Caenorhabditis* Genetics Center (CGC). N2 and HLH-30::GFP strains were cultured at 20 °C. *mel-26* mutants were maintained at 15 °C. For lifespan analysis, temperature-sensitive *mel-26* mutants and the corresponding controls were cultured at 25 °C in the presence of fluorodeoxyuridine (FUDR). For RNAi-mediated gene knockdowns, HLH-30::GFP reporter worms were cultured at 20 °C.

Generation of KLHL19 or KLHL22 knock-out cells. CRISPR guide sequences targeting the second exon of *KLHL19* or *KLHL22* were designed by http://crispr.mit.edu and cloned into pBC2 CRISPR vectors that were provided by Y. Wang (Peking University). Sequences were as follows: sgKLHL19_1_human: 5′ TTGGCATCATGAACGAGCTG 3′; sgKLHL19_2_human: 5′ GGATGCACC GGCCGCCCAGT 3′; sgKLHL22_1_human: 5′ CACTGCGTGAACAACACCTA 3′; sgKLHL22_2_human: 5′ GGACAGCGGAATCCTCTTCG 3′; sgDEP DC5_1_human: 5′ GTGTTCCCTCACATCAAGCT 3′; sgDEPDC5_2_human: 5′ AGGATCAGTATATTGGCCGT 3′; sgKLHL19_1_mouse: 5′ TCGCAGGACGG TAACCGAAC 3′; sgKLHL19_2_mouse: 5′ GGGACGCAGTGATGTATGCC 3′; sgKLHL22_1_mouse: 5′ ATCGGATTCTGCTAGCTGCA 3′; sgKLHL22_2_ mouse: 5′ GATCCTCTTTGACGTTGTCC 3′·HEK293T, MEF, MDA-MB-231 or MDA-MB-468 cells were cultured in 6-well plates and transfected with the corresponding CRISPR vector. Cells were trypsinized 48 h later and 250 μg/ml hygromycin was added into the culturing medium. Hygromycin-resistant cells were then sorted into 96-well plates and validated by genotyping.

Generation of stable cell lines. HEK293T stable cell lines were generated using a lentiviral system. The following lentiviral expression constructs were used: pBOBE-HA-DEPDC5, pBOBE-Myc-Ub, pBOBE-Myc-K48Ub, pBOBE-FLAG-KLHL22 and pLJM1-LAMP1-mRFP-FLAG$^{\times 2}$. The pBOBE vector was a gift from S.-C. Lin (Xiamen University). HEK293T cells in 6-well plates were transfected with the plasmids indicated above, together with pMDLg/pRRE, pRSV-Rev and pCI-VSVG. Medium was changed 8 h after transfection. Cells were then cultured for an additional 24 h to allow virus production (virus packaging). HEK293T cells were cultured in virus-containing medium supplemented with 8 μg/ml polybrene for 24 h (virus infection). G418 was added to the culturing medium for selection.

Immunoprecipitation. Cells were lysed on ice for 40 min using TX-100 lysis buffer (20 mM Tris-HCl, 150 mM NaCl, 1 mM Na_2EDTA, 1 mM EGTA, 2.5 mM sodium pyrophosphate, 1 mM β-glycerophosphate, 1% Triton X-100, pH 7.4) supplemented with EDTA-free protease inhibitors (Roche). Cell lysates were then centrifuged at 12,000 r.p.m. for 10 min at 4 °C. For anti-FLAG and anti-HA immunoprecipitation, supernatants of cell lysates were supplemented with washed FLAG or HA affinity gels (Sigma/Pierce) and rotated at 4 °C for 2 h. Immunoprecipitates were then washed three times with TX-100 lysis buffer. Whole cell lysates and immunoprecipitates were analysed by immunoblotting.

For transfection-based experiments, HEK293T cells were plated in 6-cm dishes and transfected with pcDNA3.3-based expression vectors using polyethylenimine (PEI) transfection reagent. 48 h after transfection, cells were subjected to treatments and then processed as described above.

Ubiquitination assays. HEK293T cells with or without the expression of exogenous proteins were lysed on ice for 10 min with RIPA lysis buffer (20 mM Tris-HCl, 150 mM NaCl, 1 mM Na_2EDTA, 1 mM EGTA, 1% NP-40, 1% sodium deoxycholate, 2.5 mM sodium pyrophosphate, 1 mM β-glycerophosphate) containing 1% SDS supplemented with EDTA-free protease inhibitors (Roche). Cell lysates were briefly sonicated, boiled at 100 °C for 10 min, and then centrifuged at 12,000 r.p.m. for 10 min at 4 °C. Supernatants were then diluted 1:3 with RIPA lysis buffer to reduce the concentration of SDS. For ubiquitination assay of HA-tagged DEPDC5, pre-washed HA affinity gels were incubated with diluted lysates for 2 h at 4 °C. For ubiquitination assay of endogenous DEPDC5, anti-DEPDC5 antibodies were added into diluted lysates for overnight binding. Protein G beads were then incubated with lysates at 4 °C for 1 h to enrich anti-DEPDC5 antibodies. Immunoprecipitates were washed three times with RIPA lysis buffer containing 0.1% SDS and analysed by immunoblotting.

For knockdown experiments, siRNAs were transfected using the RNAiMAX transfection reagent (Thermo). 48 h after transfection, cells were subjected to ubiquitination assays. Sequences of siRNAs were as follows: siKLHL19_1: 5′ GGGAGTACATCTACATGCATT 3′; siKLHL19_2: 5′ GAGTGTTACGACCCAGATA 3′; siKLHL22_1: 5′ CAGGCTACGTGCACATTTA 3′; siKLHL22_2: 5′ GCTCAACAACTTCGTATAC 3′.

For in vitro ubiquitination assays, 2.5 μl GST-CUL3/RBX1 (BPS Bioscience) mixed with 2 μg FLAG–KLHL19 or FLAG–KLHL22 purified from HEK293T cells was used. The E3 ligase complex was mixed with 2 μg HA–DEPDC5 purified from HEK293T cells, 15 μg wild-type ubiquitin or K48R ubiquitin, 550 ng E1 (UBE1), 850 ng E2 (UBE2D2), 12.5 mM Mg-ATP and 1 × ubiquitin reaction buffer, and incubated at 37 °C for 1 h. The above reagents except for the annotated ones were purchased from Boston Biochem. Reactions were terminated by boiling at 100 °C with SDS sample buffer for 10 min and analysed by immunoblotting.

RNA extraction and quantification. Cells or worms were lysed with TRIzol Reagent. Total RNA was isolated by chloroform extraction and isopropanol precipitation. One microgram total RNA was used for reverse transcription using a cDNA synthesis kit (TransGen Biotech). cDNAs of each gene were quantified by PCR or real-time PCR (qPCR). For PCR, amplifications of cDNAs were resolved by agarose gel. For qPCR, quantifications of transcripts were normalized to GAPDH. The following primers were used: *DEPDC5*: 5′ACCAGACTGTGACTCAAGTG and 5′ATAGGCACATGTGCTGACC; *GAPDH*: 5′ACCACAGTCCATGCCATCAC and 5′TCCACCACCC TGTTGCTGTA; *ACTβ*: 5′ATCTGGCACCACACCTTCTAC and 5′GGATA GCAACGTACATGGCTG; *KLHL19*: 5′GTGCTGTCATGTACCAGATC and 5′GGTTGAAGAACTCCTCTTGC; *KLHL22*: 5′GAGAGTGGAAGC ACTTCACTG and 5′GCGTAGATGTACCTGCCTACA; *mel-26*: 5′CGAGC TGTTACGTACATCCTG and 5′AGTCCAGATGGAGGTGGTAC; *tag-53*: 5′CGCTTTGATCACACAGTTGTC and 5′TCACATGAGCTGAGTGACCTG; *T08A11.1*: 5′TCTGTGCTCACGGTGGTATG and 5′CGAATTGAACCTGT GGAAGC; *rpl-32*: 5′AGGGAATTGATAACCGTGTCCG and 5′TAGGA CTGCATGAGGAGCATGT.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Nuclear/cytosolic fractionation. After amino acid treatment, cells in 6-well plates were lysed on ice for 10 min with 200 μl per well NP-40 lysis buffer (10 mM Tris-HCl, 150 mM NaCl, 0.05% NP-40) supplemented with EDTA-free protease inhibitors (Roche). Samples were centrifuged at 5,000 r.p.m. for 5 min at 4 °C to isolate the pellets (nuclei) and supernatants (cytosol). Nuclear and cytosolic fractions were boiled with SDS sample buffer and then analysed by immunoblotting.

Purification of recombinant proteins. Recombinant human HA–DEPDC5, FLAG–KLHL22, FLAG–KLHL19 and HA–KLHL22 were immunopurified from HEK293T cells. For purification of each protein, 10 × 15-cm dishes of HEK293T cells transfected with corresponding expression vectors were lysed on ice for 40 min using TX-100 lysis buffer supplemented with EDTA-free protease inhibitors (Roche). Cell lysates were then centrifuged at 12,000 r.p.m. for 10 min at 4 °C. Supernatants were mixed with 300 μl HA or FLAG affinity gels and rotated at 4 °C for 2 h. Immunoprecipitates were washed three times with TX-100 lysis buffer and three times with PBS, and eluted in 500 μl PBS containing 100 μg/ml HA or FLAG peptides. Purified proteins were analysed by Commassie G250 staining and immunoblotting. For storage, purified proteins supplemented with 20% (v/v) glycerol were kept at −20 °C.

Silver staining and mass spectrometry. To identify interaction proteins of DEPDC5 or KLHL22, 5 × 15-cm dishes of HEK293T cells stably expressing HA–DEPDC5 or FLAG–KLHL22 were lysed with TX-100 lysis buffer containing EDTA-free protease inhibitors (Roche). Cell lysates were centrifuged at 12,000 r.p.m. for 10 min at 4 °C. HA or FLAG affinity gels were added to supernatants for immunoprecipitation. Immunoprecipitates were denatured by boiling in SDS sample buffer and resolved by SDS–PAGE. Silver staining was carried out according to the manufacturer's instructions (Sigma, PROTSIL2). For interaction partners of HA–DEPDC5, bands of interest were cut from the gel and analysed by mass spectrometry; for FLAG–KLHL22, the whole lane was analysed.

Immunostaining. HEK293T cells were cultured on polylysine-coated glass coverslips (Corning) in 12-well plates (70,000 cells per well). After amino acid treatment, cells on coverslips were washed once with PBS, fixed with 4% paraformaldehyde in PBS, washed again with PBS and permeabilized with 0.1% Triton X-100 in PBS. Cells were then washed, and blocked in PBS containing 5% BSA for 30 min at room temperature. The coverslips were then incubated in PBS containing primary antibodies (1:200 dilutions) at 4 °C overnight. Cell were then rinsed, and incubated with secondary antibodies in PBS (1:200 dilutions) for 2 h at room temperature. After washing, coverslips were mounted on slides using a mounting buffer containing DAPI (Thermo) and imaged on Zeiss Fluorescence Microscopes (Imager M2 for Fig. 3c–e; LSM 710 confocal microscopy for other Figures).

In vitro reconstitution assay. For immunopurification of lysosomes, HEK293T cells stably expressing LAMP1–mRFP–FLAG$^{\times 2}$ were used. For each sample, cells in one 10-cm dish were pelleted through centrifugation at 1,000 r.p.m. for 3 min at room temperature. Cell pellets were washed once with fractionation buffer (50 mM KCl, 90 mM K-gluconate, 1 mM EGTA, 5 mM $MgCl_2$, 50 mM sucrose, 5 mM glucose, 20 mM HEPES, pH 7.4; 2.5 mM ATP and protease inhibitors were added immediately before use), resuspended in 0.8 ml fractionation buffer, and then mechanically broken by spraying six times through a 23 G needle attached to a 1-ml syringe. Cell fractions were spun down at 2,000*g* for 10 min at 4 °C to pellet the nuclei, yielding a post-nuclei supernatant (PNS). The PNS was adjusted to 2 ml, supplemented with 50 μl anti-FLAG affinity gels and rotated at 4 °C for 2 h to enrich lysosomes.

Immunopurified lysosomes were washed and resuspended in 300 μl fractionation buffer supplemented with 250 μM GTP and 100 μM GDP, in the presence or absence of 1 × amino acids. Lysosomes were then rotated at 650 r.p.m. on a thermomixer for 15 min at 37 °C (activation step). The reaction system was subsequently supplemented with 20 μl purified HA–KLHL22 and rotated for additional 25 min (binding step). Immunoprecipitates were washed, denatured and then analysed by immunoblotting.

In vitro binding assay. HEK293T cells in 10-cm dishes were transfected with vectors expressing HA–DEPDC5, HA–DEPDC5-5KR or HA–DEPDC7. 48 h after transfection, cells were lysed in 2 ml ice-cold TX-100 lysis buffer containing EDTA-free protease inhibitors (Roche). Cell lysates were centrifuged at 12,000 r.p.m. for 10 min at 4 °C. Supernatants were saved, supplemented with 30 μl anti-HA affinity beads, and rotated at 4 °C for 2 h. Immobilized HA–DEPDC5, HA–DEPDC5-5KR or HA–DEPDC7 was washed once in TX-100 lysis buffer, followed by three washes in TX-100 lysis buffer supplemented with 500 mM NaCl. TX-100 lysis buffer was then replaced with 88 μl binding buffer (40 mM HEPES, 2 mM EGTA, 2.5 mM $MgCl_2$, 0.3% CHAPS, pH 7.4). For in vitro binding assay, 88 μl immobilized HA–DEPDC5, HA–DEPDC5-5KR or HA–DEPDC7 was incubated with 12 μl immunopurified FLAG–KLHL22 and supplemented with 1% BSA and 2 mM DTT at 4 °C for 1 h. Samples were washed three times in binding buffer and subjected to immunoblotting.

Determination of cell size. Cells at the confluence of ~4×10^6 per ml were subjected to cell diameter determination using an easy cell analyser (CountStar).

HLH-30::GFP nuclear translocation assay. HLH-30::GFP reporter strain was crossed with *mel-26* mutants to generate *mel-26*; HLH-30::GFP. HLH-30::GFP or *mel-26*; HLH-30::GFP reporter strains subjected to the indicated treatments were photographed using a Zeiss Imager M2 microscope. Representative images are shown.

RNAi in *C. elegans*. RNAi clones from the Ahringer Library were grown in LB containing 50 μg/ml carbenicillin at 37 °C overnight and then seeded onto worm plates with IPTG. Dried plates were kept at room temperature overnight to allow IPTG induction of dsRNA expression. Worms were then seeded onto RNAi plates. Knockdown efficiencies were measured using qPCR.

Lifespan analysis. To compare the lifespan of *mel-26* mutants with that of N2 worms, about 100 L4 stage (day 0) N2 or *mel-26* mutant worms were transferred to plates containing *E. coli* OP50 and FUDR to prevent reproduction, and cultured at 25 °C. For lifespan analysis of *mel-26* mutants with *T08A11.1* knockdown, about 100 L4 stage *mel-26* worms were transferred to plates containing bacteria HT115 expressing dsRNAs, and cultured at 25 °C. Worms were transferred to new plates and counted every other day. FUDR was omitted after day 12. Animals that did not move when gently prodded were scored as dead. Animals that crawled off the plate or died owing to vulva bursting were not included. Sample size was determined by reference to the literature. Worms were chosen in an unbiased fashion for experimental analysis to ensure randomization. The investigators who collected the data were blinded to genotype and treatment.

Anchorage-independent growth assays. For the bottom-layer agarose medium, a 1:1 mixture of 1% low melting point agarose (Sigma) and pre-warmed 2 × culture medium (prepared from powdered DMEM or L15 medium, Gibco) was added to 6-well plates (1.5 ml per well) and solidified at 20 °C for 30 min. The 1:1 mixed 0.6% low melting point agarose and cells in 1 × culture medium with or without 100 nM rapamycin were poured onto the solidified bottom medium and allowed to solidify at 20 °C for 30 min, forming the top-layer agarose medium. Concentrations of HEK293T cells and MDA-MB-231/468 cells in 1 × medium were 6,500 cells per ml and 2,500 cells per ml. One-hundred microlitres per well of 1 × culture medium was added twice weekly onto the top medium to prevent drying. Cells were cultured at 37 °C with 5% CO_2 for around two weeks for adequate colony formation. Colonies were stained with 0.005% crystal violet in 5% methanol and quantified using ImageJ.

Tumour transplantations. Trypsinized MDA-MB-231 or MDA-MB-468 cells were washed twice with PBS and concentrated to 10^6 per 100 μl in PBS. Cell suspensions were then mixed with equal volumes of Matrigel (Corning). Two-hundred microlitres of cells mixed with matrigel were then loaded into a 1-ml insulin syringe (BD) and subcutaneously injected into the right back flank of an 8-week-old female BALB/c nude mouse. Rapamycin (LC Laboratories) in ethanol at 10 mg/ml was diluted in 5% Tween-80 (Sigma) and 5% PEG-400 (sigma). Treatment was conducted by intraperitoneal injection of 1.5 mg (low dose) or 4.5 mg (high dose) rapamycin every other day, starting at day 0 (the day of transplantation) or day 6. Xenografts were measured with a caliper every other day after transplantation (tumour volume = width^2 × length × 0.523). Tumour size must not exceed 20 mm at the largest diameter in an adult mouse, according to the IACUC and IRB of Beijing Laboratory Animal Research Center (BLARC). None of the experiments exceeded this limit in our study. Mice were killed when tumours reached 15 mm at the largest diameter or ulceration was evident. BALB/c nude mice were purchased from Charles River Laboratories and kept in a specific-pathogen-free facility at BLARC. Animals were maintained in accordance with institutional guidelines. We complied with all relevant ethical regulations of IACUA. Sample size was determined according to literatures. Mice were chosen in an unbiased fashion for experimental analysis to ensure randomization. The technicians who collected the data were blinded to genotype and treatment.

Patient samples. Breast cancer patient samples were all from Peking University Cancer Hospital, which owns a large breast cancer Biobank. Fresh breast tumour samples were obtained from patients and stored at −80 °C for further use. The tumours were all stage I–III. Tumour stage was classified according to the tumour-node-metastasis (TNM) classification of the Union International Cancer Control. This study was conducted in accordance with the ethics principles of the Declaration of Helsinki and approved by the Research and Ethics Committee of Peking University Cancer Hospital. All patients provided written informed consent.

Statistics and reproducibility. Unless otherwise stated, no statistical methods were used to predetermine sample size, the experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment. The epidemiological data for association between transcript levels of KLHL22 and each cancer type, and the related statistical significances, were provided by the Oncomine platform. For graphs with error bars or statistical significance, details of reproducibility and statistics are indicated in the corresponding figure legends. All statistical analyses in our study were conducted using GraphPad Prism 5. For other graphs showing representative data, reproducibility are stated

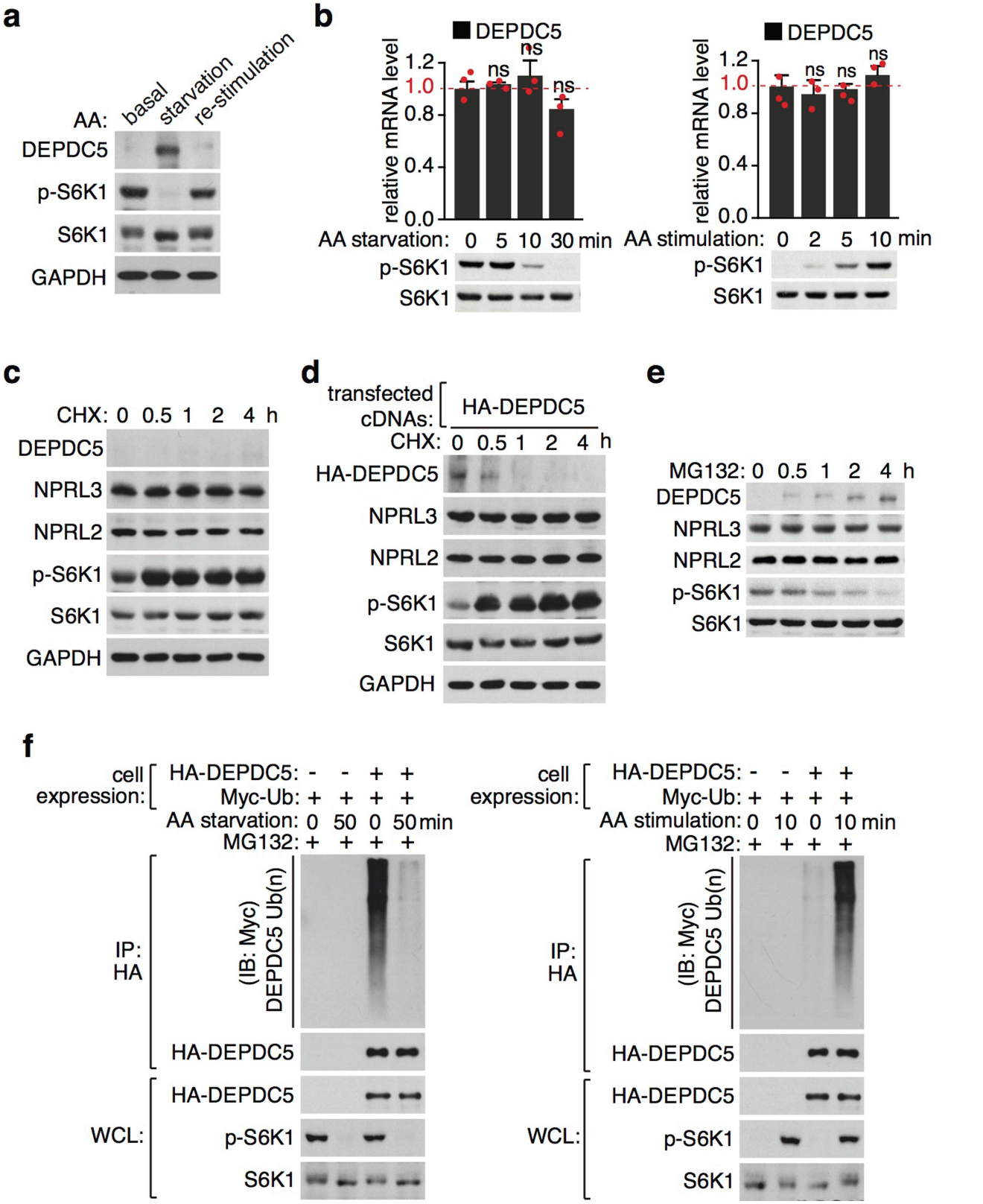
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

below: (1) $n \geq 3$ biologically independent experiments for: Figs. 1a–j; 2a–e, g (statistical analysis in Extended Data Fig. 6f, h); 3a–e; 4a, d, e; Extended Data Figs. 1a, b (immunoblotting), c–f; 2a, d; 3b–e; 4a–d, f–h; 5a, b, d, f–h; 6b–d; 8a–c, e, g, h (immunoblotting), i; 9c–e. (2) $n = 3$ technically independent experiments for Fig. 4c. (3) $n = 1$ experiment for Extended Data Figs. 2b, c; 5e, i.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

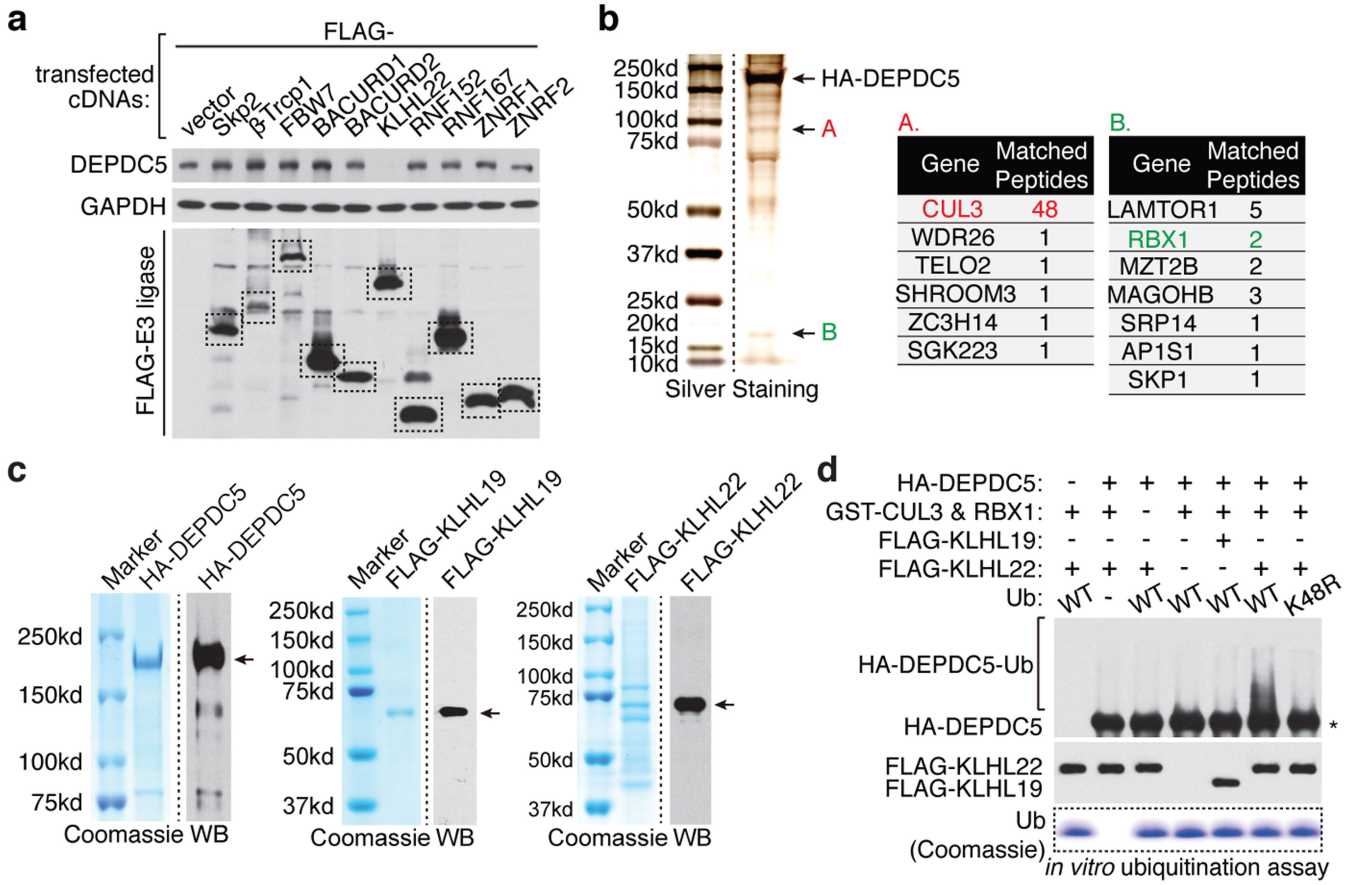
Data availability. Source Data are available in the online version of the paper. The authors declare that all data supporting the findings of this study are available within the paper and its Supplementary Information files.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



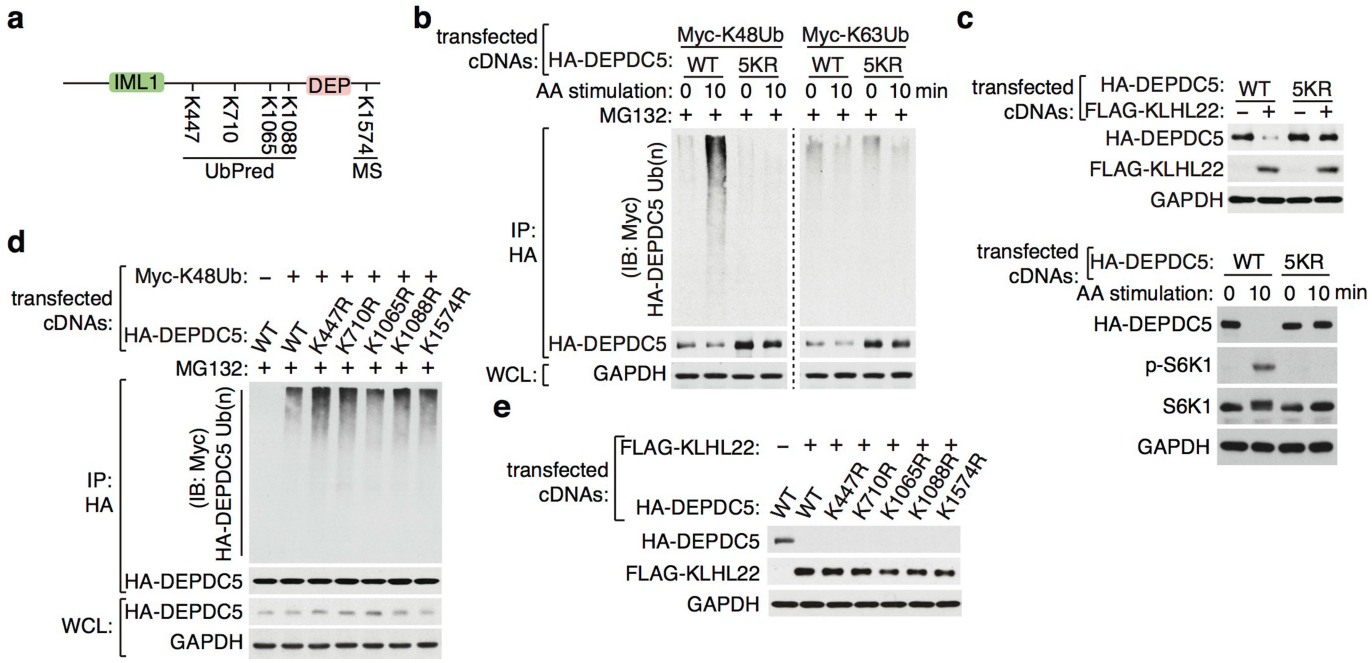
Extended Data Fig. 1 | DEPDC5 undergoes ubiquitin-mediated degradation in response to amino acids. a, Protein levels of endogenous DEPDC5 are regulated in response to the availability of amino acids. Basal, standard culture condition. **b**, Amino acids do not affect mRNA levels of DEPDC5. Relative mRNA levels of DEPDC5 to GAPDH were quantified by qPCR. $n = 3$ biologically independent experiments. ns, no significant difference; two-sided Student's *t*-test. **c**, **d**, DEPDC5 proteins are unstable under normal conditions. CHX, cycloheximide. **e**, Proteasome inhibitor MG132 increases the protein levels of DEPDC5. **f**, The ubiquitination of DEPDC5 is regulated in an amino-acid-dependent manner. For gel source data, see Supplementary Fig. 1.Source Data.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



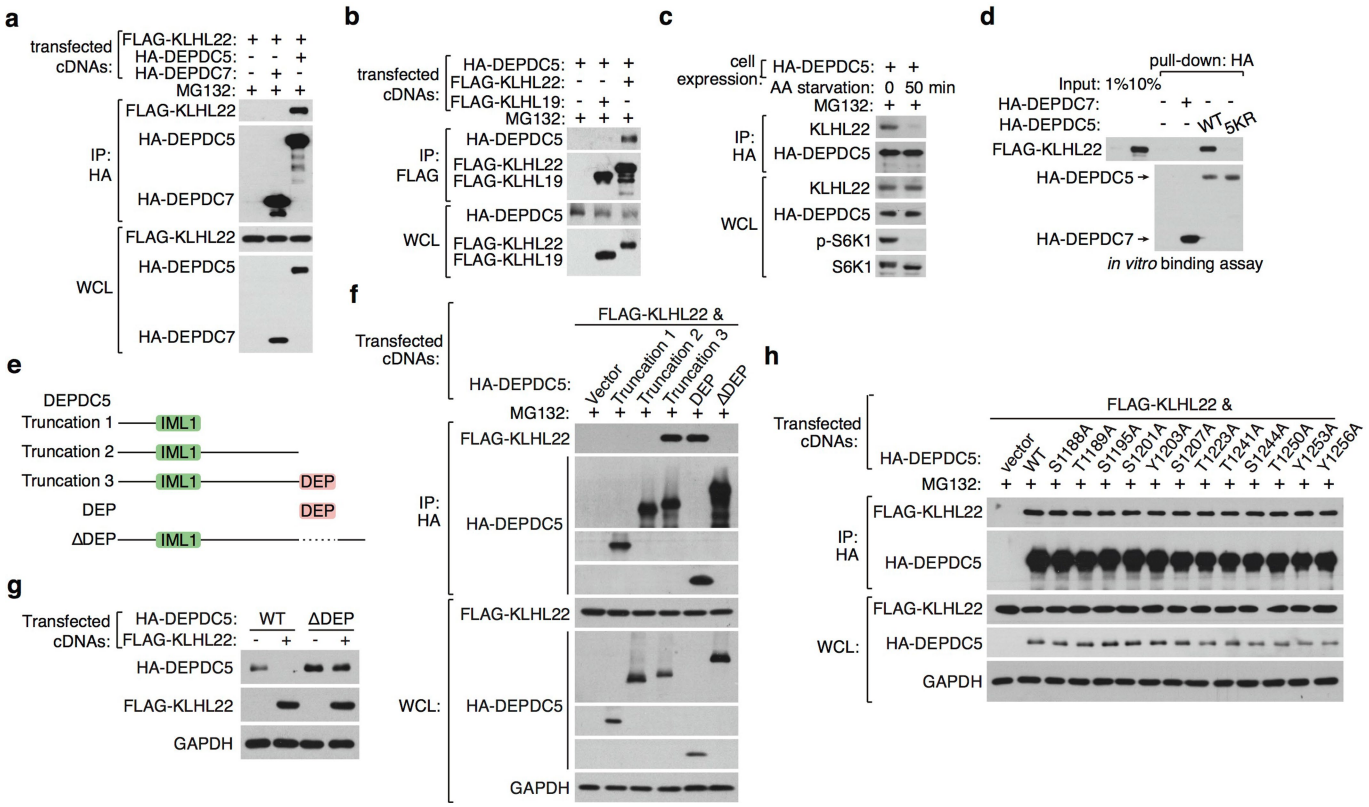
Extended Data Fig. 2 | CUL3–KLHL22 mediates the ubiquitination of DEPDC5. a, KLHL22, but not other E3 ligases, promotes the degradation of DEPDC5. **b**, DEPDC5 interacts with CUL3 and RBX1. Cells stably expressing HA–DEPDC5 were subjected to anti-HA immunoprecipitation. Arrows indicate protein bands on silver staining that were analysed by mass spectrometry. **c**, Purification of recombinant HA–DEPDC5, FLAG–KLHL19 or FLAG–KLHL22 proteins in HEK293T cells. **d**, CUL3–KLHL22 E3 ligase catalyses DEPDC5 ubiquitination in a cell-free system. Immunopurified HA–DEPDC5 and recombinant wild-type (WT) or mutant (K48R) ubiquitin were incubated with recombinant CUL3–RBX1 and KLHL22 or KLHL19. Asterisk represents HA–DEPDC5. For gel source data, see Supplementary Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



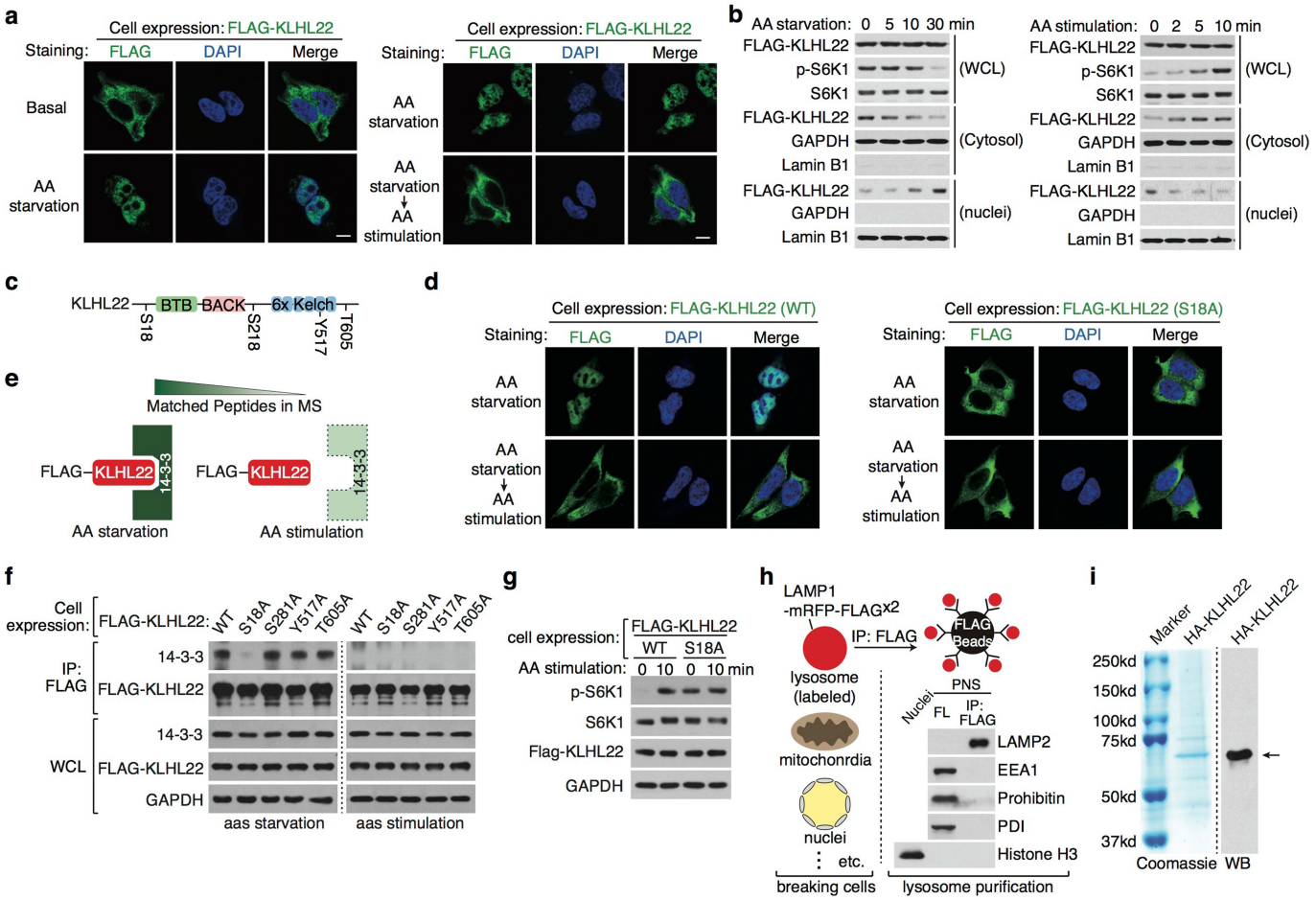
Extended Data Fig. 3 | KLHL22 promotes the ubiquitination of DEPDC5 on multiple lysine residues. a, Schematic depicting the predicted ubiquitination sites of DEPDC5. **b**, **c**, The 5KR mutation prevents K48-linked ubiquitination (**b**) and degradation (**c**) of DEPDC5. **d**, Single mutation of each lysine residue does not impair K48-linked ubiquitination (**d**) and degradation (**e**) of DEPDC5. For gel source data, see Supplementary Fig. 1.Source Data.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



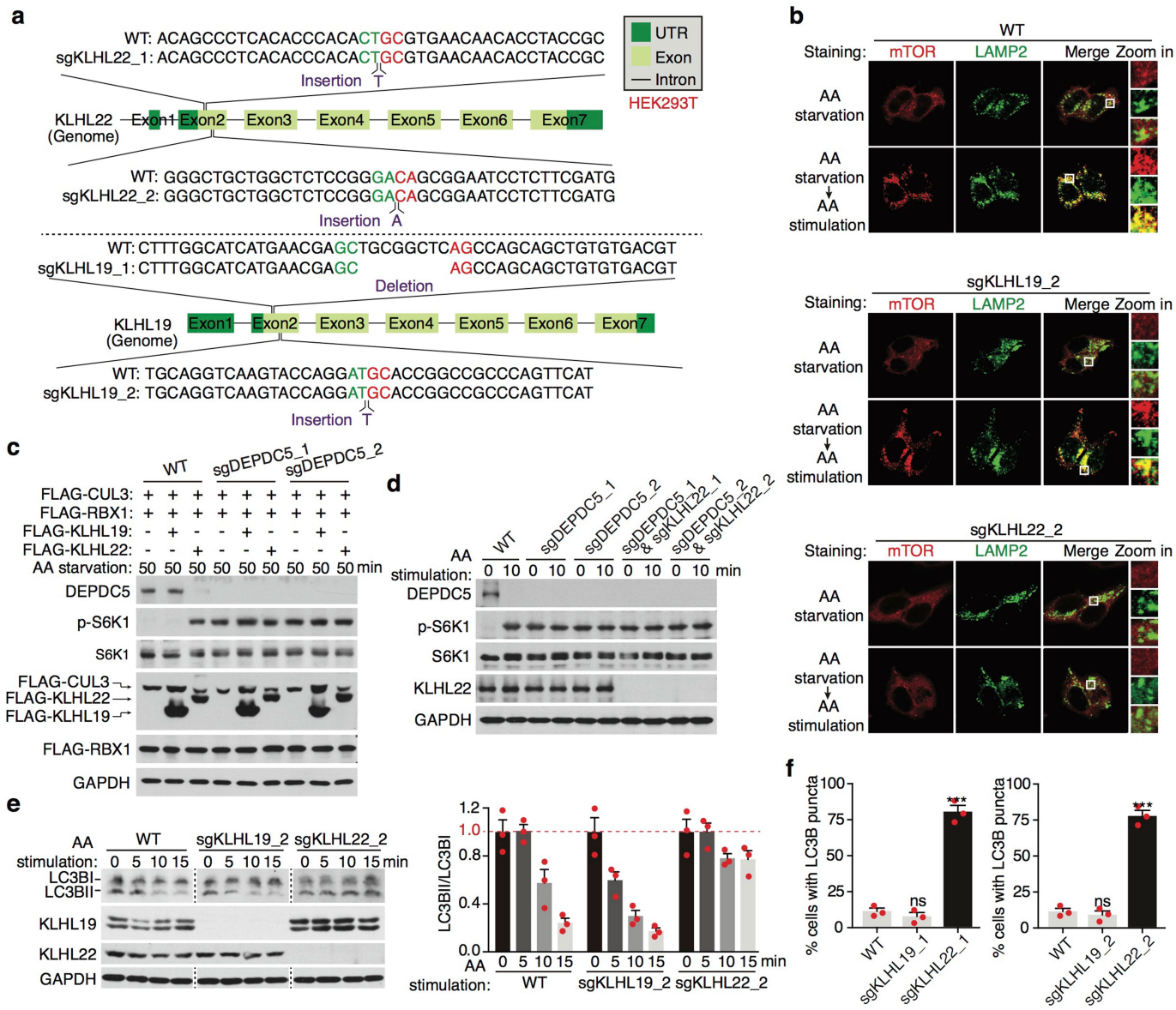
Extended Data Fig. 4 | DEP domain marks the degron of DEPDC5. **a**, **b**, KLHL22 specifically interacts with DEPDC5. **c**, KLHL22 disassociates from DEPDC5 in response to amino acid starvation. **d**, Recombinant FLAG–KLHL22 directly interacts with WT but not 5KR HA–DEPDC5 in vitro. **e**, Schematic depicting the truncations of DEPDC5. **f**, KLHL22 interacts with DEPDC5 through the DEP domain. **g**, Deletion of DEP motif prevents KLHL22-mediated degradation of DEPDC5. **h**, Mutation of each serine, threonine or tyrosine residue in DEP domain to alanine does not affect the interaction between KLHL22 and DEPDC5. For gel source data, see Supplementary Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



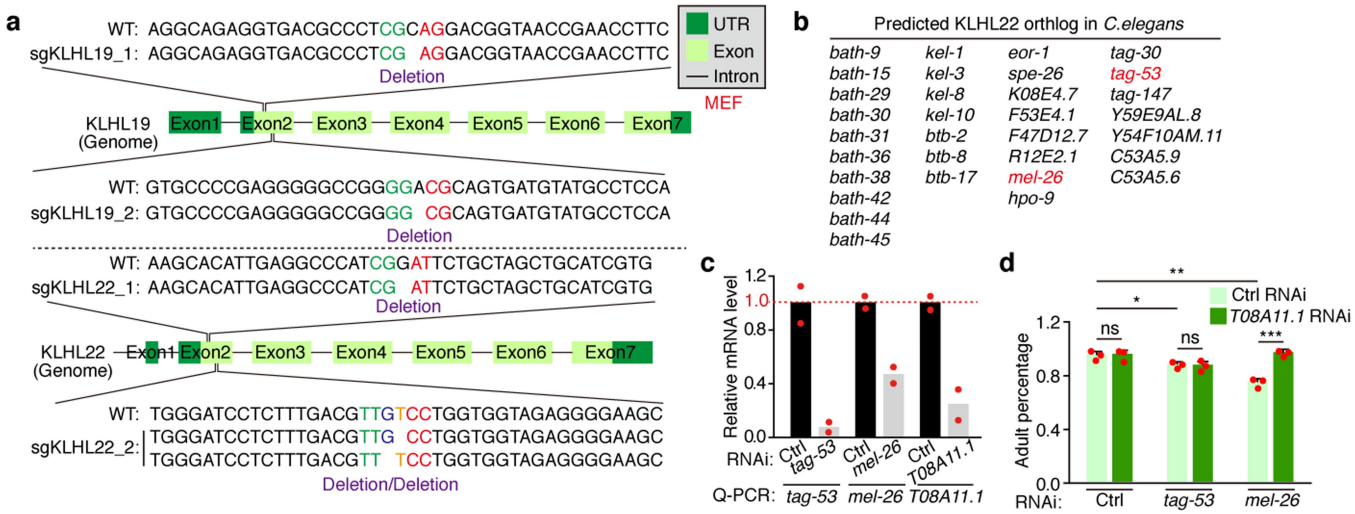
Extended Data Fig. 5 | 14–3–3 regulates nuclear–cytosolic shuttling of KLHL22. a, **b**, Nuclear–cytosolic shuttling of KLHL22 is regulated by amino acids. DAPI, nuclei. Scale bar, 10 μm. **c**, Schematic depicting the reported phosphorylation sites of KLHL22. **d**, S18A mutation prevents nuclear accumulation of KLHL22 in amino-acid-deficient conditions. Scale bar, 10 μm. **e**, Schematic depicting the interaction between FLAG–KLHL22 and 14-3-3 in mass spectrometry analysis using FLAG–KLHL22 as the bait. **f**, S18A mutation prevents the interaction between KLHL22 and 14-3-3 in amino-acid-deprived conditions. **g**, KLHL22(S18A) mutant promotes S6K1 phosphorylation in amino-acid-deficient conditions. **h**, Immunopurification of lysosomes. LAMP2 (lysosome), EEA1 (early endosome), prohibitin (mitochondria), PDI (endoplasmic reticulum) and histone H3 (nucleus). **i**, Purification of HA–KLHL22. For gel source data, see Supplementary Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



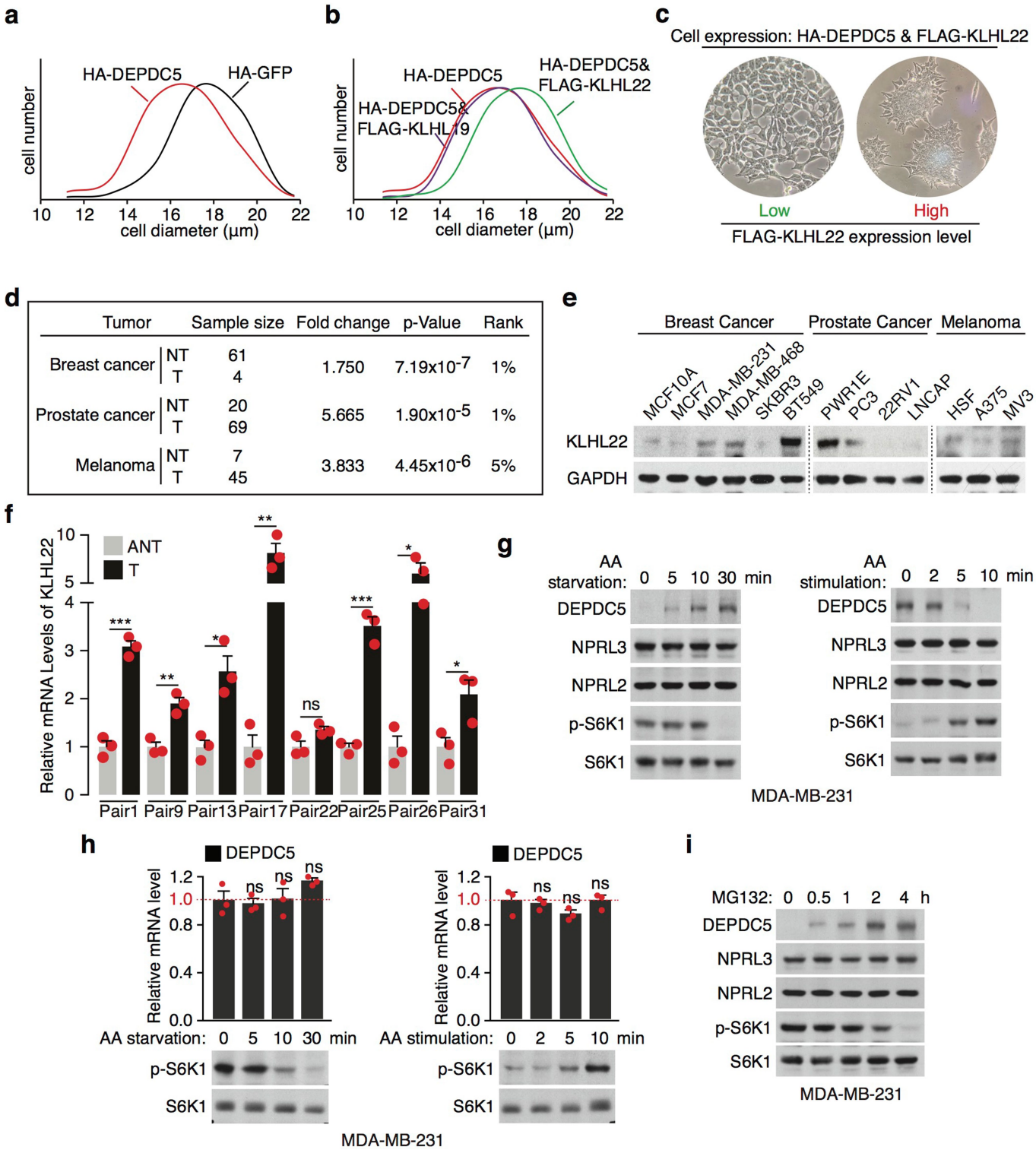
Extended Data Fig. 6 | KLHL22 is essential for activation of mTORC1 and downstream events. a, Schematic depicting CRISPR–Cas9-mediated knockout of *KLHL19* or *KLHL22* in HEK293T cells. **b**, KLHL22 is required for amino-acid-dependent lysosomal localization of mTORC1. Scale bar, 10 μm. **c**, **d**, Overexpression (**c**) or depletion (**d**) of KLHL22 does not affect mTORC1 activity in DEPDC5-deprived cells. **e**, KLHL22 is required for amino-acid-induced suppression of autophagy. $n = 3$ biologically independent experiments. **f**, Depletion of KLHL22 induces autophagy. Percentages of cells with obvious LC3B puncta were calculated. $n = 3$ biologically independent experiments. ns, no significant difference; $^{***}P < 0.001$; two-sided Student's *t*-test. For gel source data, see Supplementary Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



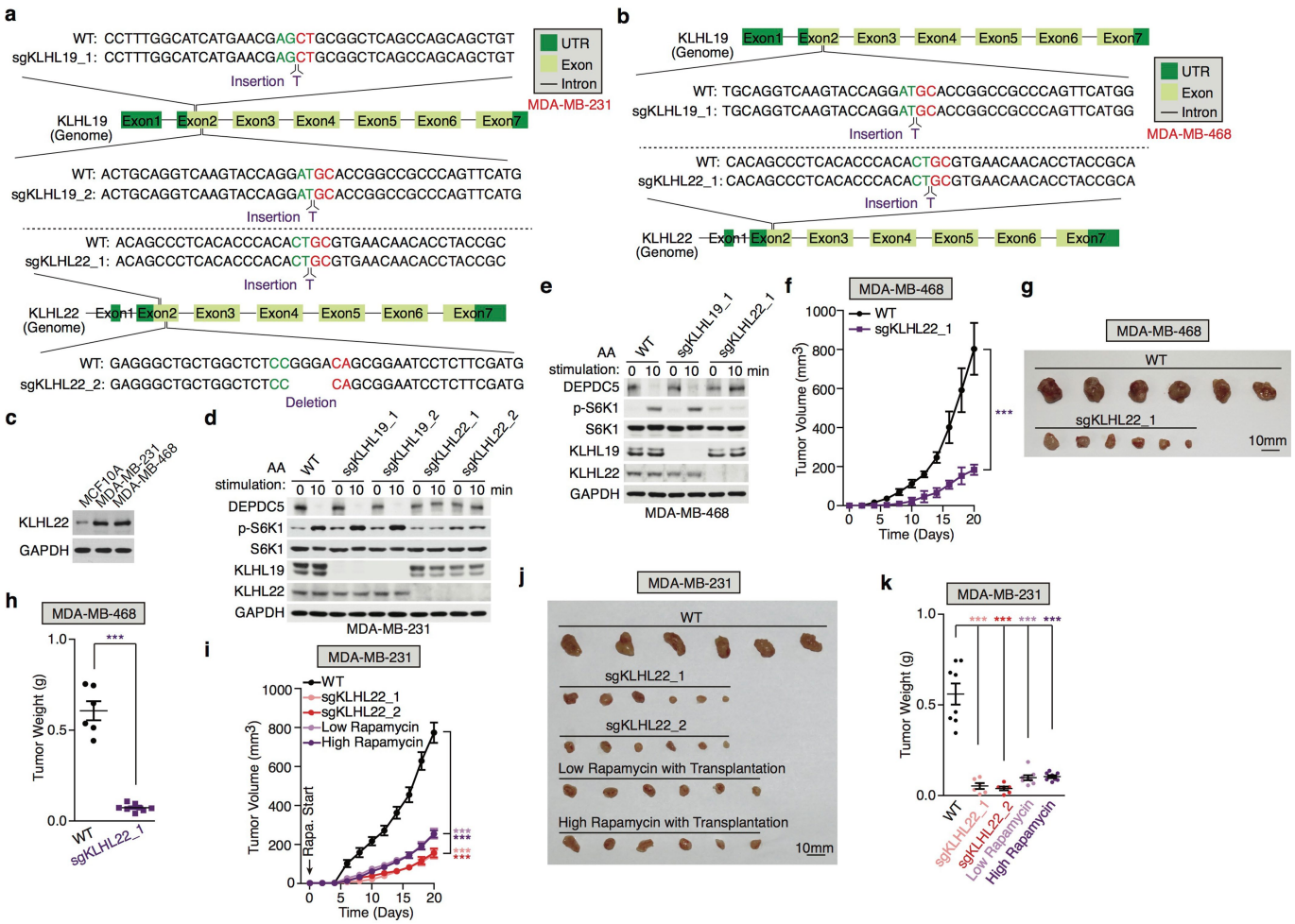
Extended Data Fig. 7 | KLHL22 has a conserved role in mice and nematodes. a, Schematic depicting CRISPR–Cas9-mediated knockout of *KLHL19* or *KLHL22* in MEF cells. **b**, Predicted orthologues of KLHL22 in *C. elegans*. **c**, Knockdown efficiency of *tag-53*, *mel-26* and *T08A11.1*. $n = 2$ biologically independent experiments. **d**, RNAi targeting *T08A11.1* suppresses the developmental delay of worms induced by *mel-26* RNAi, but not *tag-53* RNAi. $n = 3$ biologically independent experiments. ns, no significant difference; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; two-sided Student's *t*-test.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



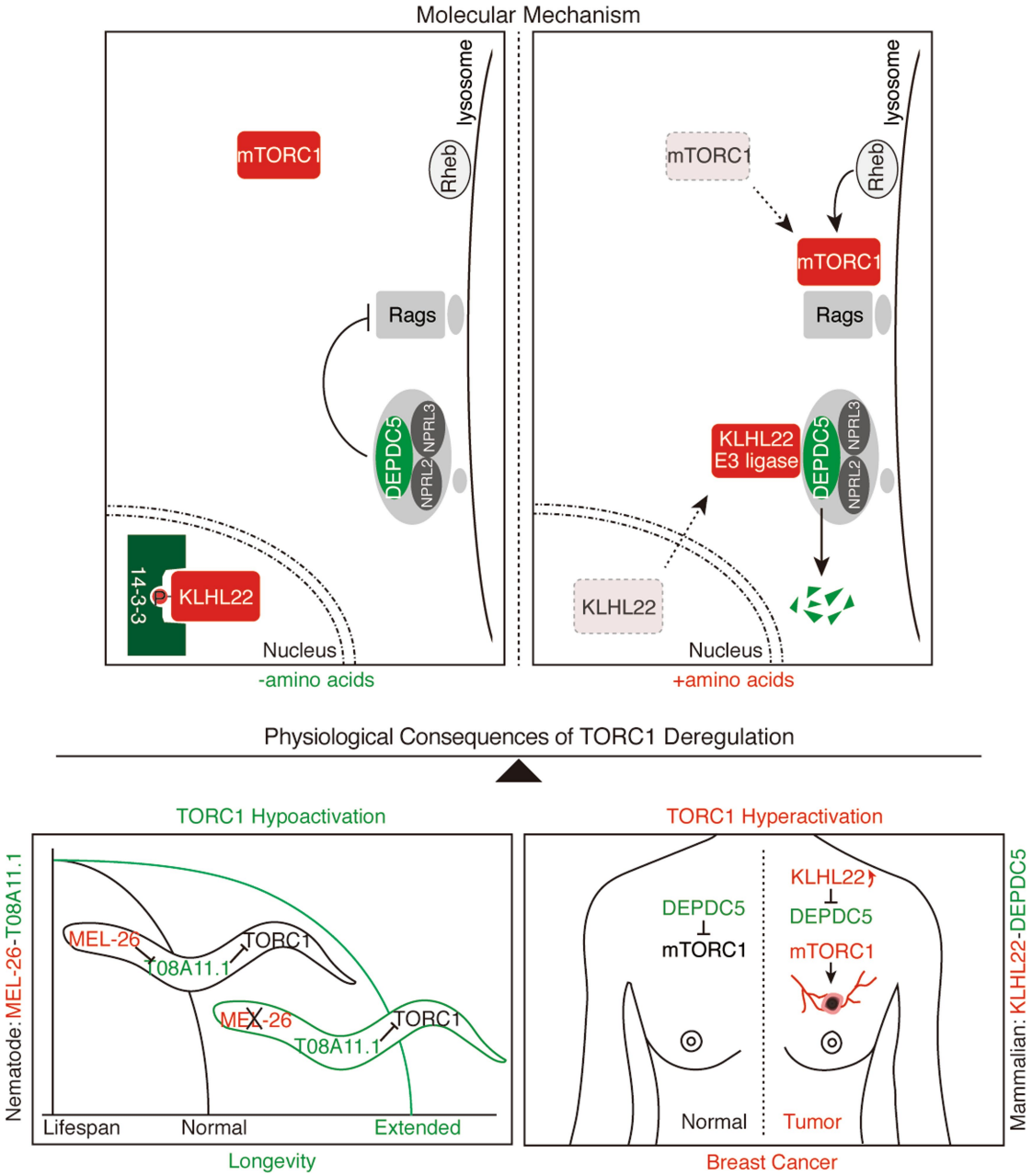
Extended Data Fig. 8 | Protein levels of KLHL22 are elevated in human breast cancer. a, Stable expression of HA–DEPDC5 reduces the size of HEK293T cells. **b**, Stable expression of FLAG–KLHL22 reverses the reduction in cell size in HEK293T cells stably expressing HA–DEPDC5. **c**, High expression of FLAG–KLHL22 transforms HEK293T cells. **d**, According to Oncomine database, mRNA levels of KLHL22 are elevated in breast and prostate cancers and in melanoma. Fold change of KLHL22 transcript levels in tumour tissues (T) to the corresponding normal tissues (NT) are shown. Rank represents KLHL22 rank in ordered list of genes that are upregulated. **e**, Protein levels of KLHL22 in breast cancer, prostate cancer and melanoma cell lines and corresponding normal cells (MAF10A, PWR1E and HSF). **f**, mRNA levels of KLHL22 are elevated in tumours of breast cancer patients. T: tumours; ANT: adjacent normal tissues. Quantifications are normalized to actin. $n = 3$ technically independent experiments. ns, no significant difference; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; two-sided Student's *t*-test. **g**, Protein levels of endogenous DEPDC5 are regulated in an amino-acid-sensitive manner in MDA-MB-231 cells. **h**, Amino acid deprivation or stimulation does not affect mRNA levels of DEPDC5 in MDA-MB-231 cells. Relative mRNA levels of DEPDC5 to GAPDH were quantified using qPCR. $n = 3$ biologically independent experiments. ns, no significant difference; two-sided Student's *t*-test. **i**, Proteasome inhibitor MG132 increases the protein levels of DEPDC5 in MDA-MB-231 cells. For gel source data, see Supplementary Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | Deletion of KLHL22 in breast cancer cells prevents mTORC1 activation and tumorigenesis. a, **b**, Schematic of CRISPR–Cas9-mediated knockout of *KLHL19* or *KLHL22* in MDA-MB-231 (**a**) or MDA-MB-468 (**b**) breast cancer cells. **c**, Protein levels of KLHL22 are upregulated in MDA-MB-231/468 cells. **d**, **e**, Deletion of KLHL22 in MDA-MB-231 (**d**) or MDA-MB-468 (**e**) cells suppresses amino-acid-induced DEPDC5 degradation and S6K1 phosphorylation. **f**–**h**, Deletion of KLHL22 suppresses tumour growth of MDA-MB-468 cells. Tumour volume, mean ± s.e.m. ($n = 9$ mice for WT, $n = 8$ mice for sgKLHL22_1), $^{***}P < 0.001$, two-sided ANOVA (**f**); tumour images (**g**); tumour weights 20 days after transplantation, mean ± s.e.m. ($n = 6$ mice for WT and $n = 8$ mice for KLHL22 knockouts), $^{***}P < 0.001$, two-sided Student's *t*-test (**h**). **i**–**k**, Deletion of KLHL22 or treatment with rapamycin suppresses tumour growth of MDA-MB-231 cells. Tumour volume, mean ± s.e.m. ($n = 9$ mice for sgKLHL22_2 and WT injected with low dose of rapamycin, $n = 10$ mice for remaining groups), $^{***}P < 0.001$, two-sided ANOVA (**i**); tumour images (**j**); and tumour weights 20 days after transplantation, mean ± s.e.m. ($n = 8$ mice for WT and WT injected with low or high dose of rapamycin; $n = 6$ mice for KLHL22 knockouts), $^{***}P < 0.001$, two-sided Student's *t*-test (**k**). For gel source data, see Supplementary Fig. 1.Source Data.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Proposed model of KLHL22-mediated regulation of mTORC1, ageing and cancer. In response to amino acids, KLHL22 translocates from the nucleus to the cytosol, where it accumulates on the surface of lysosomes to mediate the ubiquitination and degradation of DEPDC5, an essential subunit of GATOR1. GATOR1 loss of function activates Rag GTPases and subsequently activates mTORC1 (top). KLHL22-mediated DEPDC5 degradation is a conserved mechanism of TORC1 regulation in mammals and nematode. TORC1 hypoactivation due to *mel-26* (*Ce*.KLHL22) depletion extends *C. elegans* lifespan, whereas TORC1 hyperactivation due to elevated expression of KLHL22 promotes human breast cancer (bottom).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0120-4

Structural basis for gating pore current in periodic paralysis

Daohua Jiang[1,6], Tamer M. Gamal El-Din[1,6], Christopher Ing[2,3], Peilong Lu[1,4], Régis Pomès[2,3], Ning Zheng[1,5]* & William A. Catterall[1]*

Potassium-sensitive hypokalaemic and normokalaemic periodic paralysis are inherited skeletal muscle diseases characterized by episodes of flaccid muscle weakness[1,2]. They are caused by single mutations in positively charged residues ('gating charges') in the S4 transmembrane segment of the voltage sensor of the voltage-gated sodium channel $Na_v1.4$ or the calcium channel $Ca_v1.1$[1,2]. Mutations of the outermost gating charges (R1 and R2) cause hypokalaemic periodic paralysis[1,2] by creating a pathogenic gating pore in the voltage sensor through which cations leak in the resting state[3,4]. Mutations of the third gating charge (R3) cause normokalaemic periodic paralysis[5] owing to cation leak in both activated and inactivated states[6]. Here we present high-resolution structures of the model bacterial sodium channel Na_vAb with the analogous gating-charge mutations[7,8], which have similar functional effects as in the human channels. The R2G and R3G mutations have no effect on the backbone structures of the voltage sensor, but they create an aqueous cavity near the hydrophobic constriction site that controls gating charge movement through the voltage sensor. The R3G mutation extends the extracellular aqueous cleft through the entire length of the activated voltage sensor, creating an aqueous path through the membrane. Conversely, molecular modelling shows that the R2G mutation creates a continuous aqueous path through the membrane only in the resting state. Crystal structures of Na_vAb(R2G) in complex with guanidinium define a potential drug target site. Molecular dynamics simulations illustrate the mechanism of Na^+ permeation through the mutant gating pore in concert with conformational fluctuations of the gating charge R4. Our results reveal pathogenic mechanisms of periodic paralysis at the atomic level and suggest designs of drugs that may prevent ionic leak and provide symptomatic relief from hypokalaemic and normokalaemic periodic paralysis.

$Na_v1.4$ channels generate action potentials that initiate muscle contraction[9]. They are complexes of a pore-forming α-subunit and auxiliary β1 subunits[9–11]. The α-subunit contains four homologous domains (I–IV), each with six transmembrane segments (S1–S6). Segments S1–S4 form the voltage sensor, and every third residue in S4 is positively charged. Upon depolarization, S4 moves outward through a narrow gating pore formed by S1–S3, catalysed by interactions with negative or polar residues in S2 and S3[12]. The voltage sensor has an hourglass shape, with a narrow hydrophobic constriction site (HCS) that separates extracellular and intracellular compartments[6,11]. Water-filled crevices on either side of the HCS focus the membrane electric field, assuring efficient coupling of voltage to conformational changes that open the central pore[12,13]. Mutations in the arginine gating charges that occupy the HCS cause state-dependent cation leak through the voltage sensor, which we term 'gating pore current'[14,15].

Missense mutations of arginine gating charges in S4 of $Na_v1.4$ cause hypokalaemic periodic paralysis and normokalaemic periodic paralysis[1,2,16,17]. Mutations of R1 in domains I or III to H or Q, or mutation of R2 in domains I, II and III to W, G, Q or S cause hypokalaemic periodic paralysis[1,2,16,17]. Mutations of R3 in domain II to G, Q or W, or of R3 in domain III to H or C cause normokalaemic periodic paralysis[4,16,17]. All these mutations result in non-selective gating pore current through the voltage sensor[3,4,16–19]. Increased inward leak leads to Na^+ overload, sustained depolarization and action potential failure, which paralyze skeletal muscles[3,16–19]. These pathophysiological effects suggest that mutations that cause hypokalaemic periodic paralysis result in an open aqueous pathway for ion movement in the resting state of the voltage sensor, but not in the activated state, and mutations that cause normokalaemic periodic paralysis result in an open aqueous pathway in the activated state, but not in the resting state. Molecular models and mutagenesis studies support this hypothesis[20–22]. To provide direct structural evidence for this pathophysiological mechanism, we introduced mutations known to cause periodic paralysis into Na_vAb, a voltage-gated Na^+ channel from *Arcobacter butzleri*, the structure of which has been solved at high resolution[6,7]. We characterized the resulting gating pore currents, solved the structures of mutant gating pores without and with a bound permeant ion, and investigated molecular dynamics[23] of ion movement through the gating pores.

To reconstitute pathogenic hypokalaemic periodic paralysis gating pore currents in Na_vAb, we mutated R2 to S (R2S, analogous to $Na_v1.4$(R672S)) and expressed the mutant in *Trichopulsia ni* insect cells. Transfected cells were voltage-clamped to −200 mV and depolarized in 10-mV steps to record Na^+ currents. Half-maximal activation of central pore currents was observed at $V_a = -105 \pm 0.6$ mV (Fig. 1a). To measure gating pore currents, cells were held at −100 mV, at which Na_vAb is in the slow-inactivated state and exhibits no central pore current. Gating pore current was examined by applying pulses from + 100 to −200 mV in −10 mV steps. A nonlinear leak current component was observed in the resting state, beginning at −110 mV and increasing to −200 mV (Fig. 1b, c).

Mutations of the gating charge R3 that cause normokalaemic periodic paralysis ($Na_v1.4$(R675G/Q/W)) induce outward gating pore current in activated but not in resting states[6]. In Na_vAb(R3G), central pore current was activated between −50 mV and 0 mV (Fig. 1d; $V_a = -24.8 \pm 1.1$ mV). Steady-state inactivation was observed from −90 mV to −10 mV with half maximal inactivation at $V_h = -47.7 \pm 0.4$ mV (Fig. 1d). Na_vAb(R3G) conducted outward gating pore current in both activated and inactivated states at potentials more positive than −60 mV (Fig. 1e, f). These physiological studies demonstrate that Na_vAb provides an accurate model of $Na_v1.4$, because gating pore current is observed only in the resting state for Na_vAb(R2S) and only in the activated and inactivated states for Na_vAb(R3G).

The pathogenic effects of gating pore mutations depend on inward leak of Na^+. The R2S mutant gating pore was not significantly selective among Cs^+, K^+ or Na^+ (Fig. 1g, $P > 0.7$). As is the case for $Na_v1.4$[24], the gating pore of Na_vAb(R2S) was exceptionally permeant to guanidinium (about 28-fold greater than Na^+), but it was less permeant to methylguanidinium and ethylguanidinium (Fig. 1g). The outward gating pore currents conducted by Na_vAb(R3G) were higher for Cs^+ than for K^+ or Na^+, which were similar to each other (Fig. 1h). However, Na_vAb(R3G)

[1]Department of Pharmacology, University of Washington, Seattle, WA, USA. [2]Molecular Medicine, Hospital for Sick Children Toronto, Toronto, Ontario, Canada. [3]Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada. [4]Institute for Protein Design, University of Washington, Seattle, WA, USA. [5]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [6]These authors contributed equally: Daohua Jiang, Tamer M. Gamal El-Din. *e-mail: wcatt@uw.edu; nzheng@uw.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

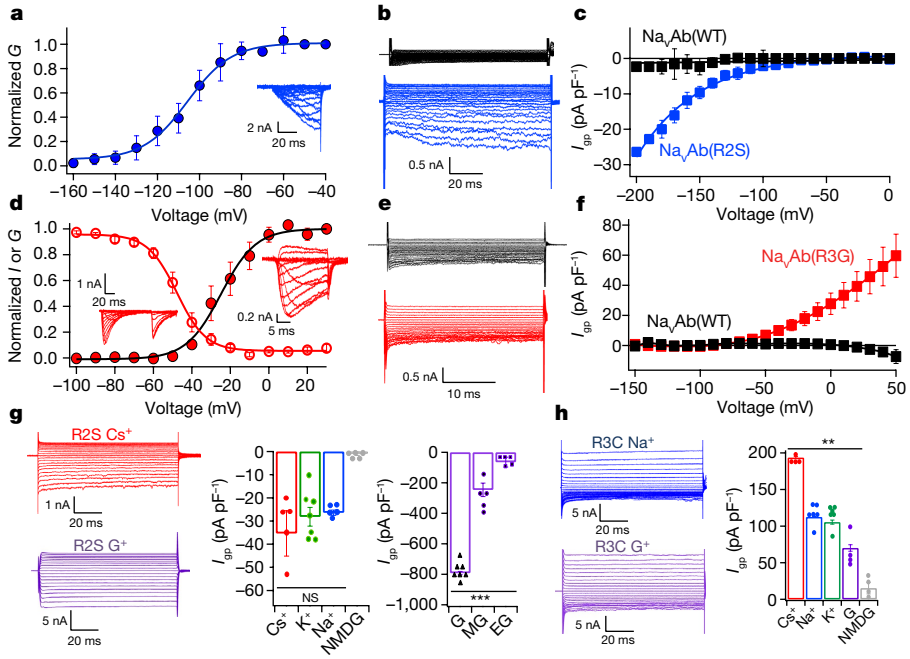


Fig. 1 | Functional properties of Na_vAb(WT), Na_vAb(R2S) and Na_vAb(R3G). a, Central pore Na^+ currents (inset) and conductance–voltage (G–V) curve for Na_vAb(R2S) during 200-ms depolarizations from −200 mV to the indicated potentials. V_a (the voltage for half-maximal activation) = −105 ± 0.6 mV; slope factor, k = 10 ± 0.9. n = 4. **b**, **c**, Gating pore Na^+ currents (I_{gp}) and current–voltage (I–V) curves for Na_vAb(R2S) (blue) or Na_vAb(WT) (black) during test pulses from −100 mV to the indicated potentials. n = 10. **d**, Central pore Na^+ currents (inset) and G–V curve for Na_vAb(R3G) during depolarizations from −160 mV to the indicated potentials (filled circles; V_a = −24.8 ± 1.1 mV, k = 9 ± 1; n = 4). Voltage dependence of steady-state inactivation (open circles) for Na_vAb(R3G) (V_h = −47.7 ± 0.4 mV, k = 7.5 ± 0.3 (n = 4); V_h, the voltage for half-maximal inactivation)). **e**, **f**, Gating pore Na^+ currents and I–V curves for Na_vAb(R3G) (red) or Na_vAb(WT) (black) for voltage steps from 0 mV to the indicated potentials. n = 11. **g**, Gating pore current through Na_vAb(R2S) for Cs^+ (n = 5), K^+ (n = 7), Na^+ (n = 5), *N*-methyl-D-glucamine (NMDG, n = 5), guanidinium (G, n = 7), methylguanidinium (MG, n = 5) and ethylguanidinium (EG, n = 5) at −200 mV. ***P = 0.00029. **h**, Gating pore current through Na_vAb(R3C) for Cs^+ (n = 4), Na^+ (n = 6), K^+ (n = 6), guanidinium (n = 4) and NMDG (n = 4) at +100 mV. **P = 0.0011. Student's t-test, two-sided.

was less permeant to guanidinium than to Na^+ (Fig. 1h), and it was more than 16-fold less permeant to guanidinium than Na_vAb(R2S). The weak selectivity of R2S and R3G mutants for different inorganic cations and the high guanidinium permeability through the R2S mutant are characteristic of the corresponding mutations in $Na_v1.4$[2], further supporting the validity of Na_vAb as a model for structural studies of gating pore mutations.

To elucidate the structure of a pathogenic gating pore in its conductive conformation in an activated voltage sensor, we solved the structure of a Na_vAb analogue of a normokalaemic periodic paralysis-causing mutation, Na_vAb(R3G), at 2.7 Å resolution (Fig. 2). Voltage-gated sodium channels have a central pore module surrounded by four symmetrically located voltage sensors (Fig. 2a). The voltage sensors of Na_vAb and $Na_v1.4$ are very similar in amino acid sequence

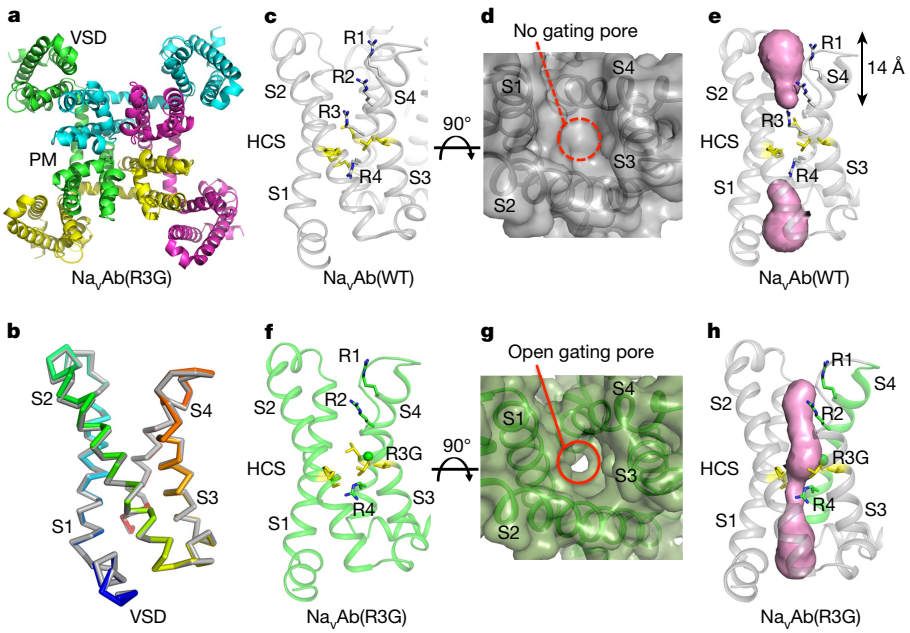


Fig. 2 | Structures of the voltage sensor of Na_vAb(WT) and Na_vAb(R3G). a, Structure of Na_vAb(R3G) in top view. **b**, Comparison of the conformations of Na_vAb(WT) (grey) and Na_vAb(R3G) (rainbow) voltage sensor in side view. **c**–**e**, Structures of Na_vAb(WT) voltage sensor. **c**, Side view highlighting gating charges in sticks. **d**, Top view in space-filling format. **e**, MOLE2 analysis of water-accessible space in magenta. **f**–**h**, Structures of Na_vAb(R3G) voltage sensor. **f**, Side view highlighting gating charges. **g**, Top view in space-filling format. **h**, MOLE2 analysis of water-filled space in magenta. Green spheres in **f** and **h** indicate the positions of the missing side chain of R3. In **d** and **g**, the dotted red line circles the position where the gating pore would be in the activated state and the solid red line circles the open gating pore, respectively. See Extended Data Table 1 for details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

and structure (Extended Data Figs. 1 and 2). The voltage sensors of wild-type (WT) Na_vAb and Na_vAb(R3G) crystallize in the same conformation, with a Cα root mean square deviation (r.m.s.d.) of 0.39 Å (Fig. 2b, Extended Data Fig. 3). These results indicate that the R3G mutation does not perturb the overall structure of the voltage sensor and, therefore, that its pathogenic effects are caused by the loss of the R3 side chain. These channels crystallize with an activated voltage sensor[6,7] (Fig. 2c), as would be expected at 0 mV. In Na_vAb(WT), R1, R2 and R3 are located extracellularly relative to the HCS, and their side chains point outward, toward the extracellular milieu (Fig. 2c). By contrast, R4 is located intracellularly relative to the HCS and its side chain points inward towards the cytosol (Fig. 2c). When viewed from the extracellular side, there is no water-accessible path into the cell through the wild-type voltage sensor (Fig. 2d); however, we observed a deep solvent-accessible cleft extending down to the R4 side chain in Na_vAb(R3G) (Fig. 2g).

Analysis of the structure of chain B of Na_vAb(WT) using the MOLE2 algorithm revealed an incomplete water-accessible path extending part of the way through the voltage sensor from both extracellular and intracellular sides, which is interrupted at the HCS by R3 (Fig. 2e). Strikingly, in Na_vAb(R3G), the water-accessible path continues all the way through the voltage sensor, and has a diameter of 2 Å at its narrowest point, similar to the size of Na^+ (Fig. 2h). By contrast, in chain A, R4 was captured in a rotamer conformation in which the arginine side chain partially blocks the inner end of the gating pore in Na_vAb(R3G) (Extended Data Fig. 4a). Previously reported structures of Na_vAb in the slow-inactivated state show that R4 adopts four slightly different rotamer conformations, with the most open having a diameter of 3 Å[25] (Extended Data Fig. 4b). These results elucidate the molecular mechanism by which mutations in S4 cause pathogenic gating pore currents and suggest that ion permeation through the gating pore is controlled dynamically by the state of the voltage sensor and by rotamer conformations of R4.

In contrast to voltage-gated sodium-channel mutations that cause normokalaemic periodic paralysis, those that cause hypokalaemic periodic paralysis result in a channel that conducts gating pore current in the resting state but is closed in the activated state (Fig. 1). Therefore, we hypothesized that Na_vAb(R2G) would not have a continuous water-accessible path through its gating pore in the activated state. Analysis of the 2.9 Å structure of Na_vAb(R2G) revealed a gap with additional solvent-accessible area in the extracellular aqueous cleft in comparison to the wild-type channel, but no change in the backbone conformation (Fig. 3a, Extended Data Fig. 3). Although the increased opening of the aqueous cleft in the voltage sensor is evident in space-filling models (Fig. 3b), the R3 and R4 side chains seal the voltage sensor in this activated state, interrupting the transmembrane path and preventing ion conductance. The solvent-accessible area penetrates about 21 Å into the membrane from the extracellular side (Fig. 3c), more than 7 Å deeper than in Na_vAb(WT) (Fig. 2e), but it does not reach the cytosolic side. This structure illustrates why Na_vAb(R2G) does not conduct gating pore current in the activated state (Fig. 1).

There are no crystal structures of the voltage sensor of a voltage-gated sodium channel in the resting state, because the resting state is only accessible at negative membrane potentials. However, we developed models of three resting states using disulfide locking of substituted cysteine residues and structure prediction with the Rosetta algorithm[26]; these are now considered consensus models of the actual resting states[27,28]. To model an open gating pore with the voltage sensor in the resting state, we introduced the R2G and R3G mutations into these resting-state models and analysed the resulting structures with the MOLE2 algorithm (Fig. 3d–f). There is no continuous path through the voltage sensor in the wild-type resting-state structure (Fig. 3d), whereas the resting state of the Na_vAb(R2G) voltage sensor contains a continuous water-accessible path through the membrane (Fig. 3e). Loss of the R2 side chain leaves a gap at the HCS that is large enough for Na^+ to pass through (Fig. 3e). By contrast, the transmembrane pathway is incomplete in Na_vAb(R3G) because the R2 side chain occupies the HCS and

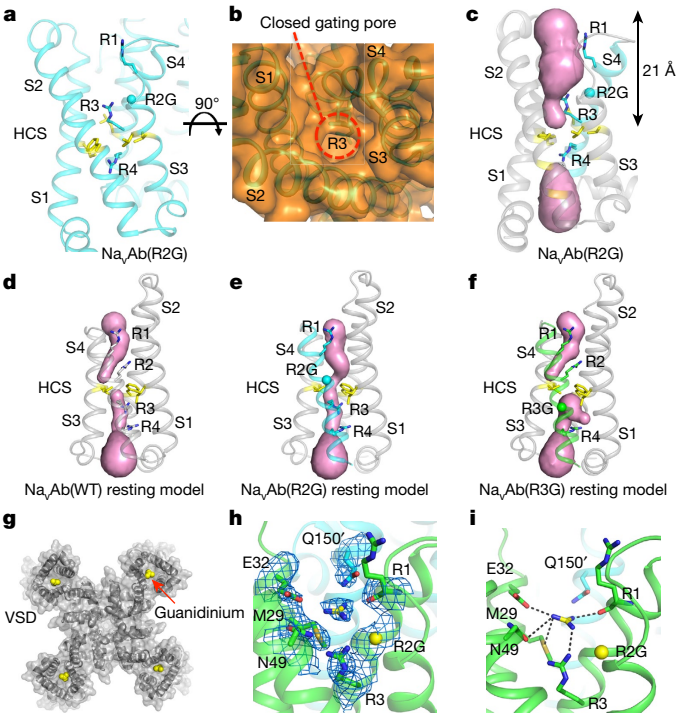


Fig. 3 | Structure of voltage sensor and guanidinium binding site of Na_vAb(R2G). a–c, Structures of the activated voltage sensor of Na_vAb(R2G). **a**, Side view with gating charges highlighted in sticks. **b**, Top view in space-filling format. The dashed red line indicates the position of the closed gating pore. **c**, MOLE2 analysis of water-filled space in magenta. **d–f**, Rosetta structural models of resting state 2 of the voltage sensor were re-optimized with the amino-acid sequence of Na_vAb for Na_vAb(WT) (**d**), Na_vAb(R2G) (**e**) and Na_vAb(R3G) (**f**). The perspective is rotated approximately 180° around the vertical axis to better illustrate the arginine gating charges in resting state 2. Green spheres represent missing arginine side chains of R2 and R3, respectively. Magenta blobs represent solvent-accessible volume modelled with MOLE2. **g**, Top view of Na_vAb(R2G) with one guanidinium bound to each voltage sensor. **h**, $2mF_o-DF_c$ electron density map (blue mesh) of residues around the guanidinium binding site at 1σ. **i**, Interaction network between guanidinium and amino acids in the voltage sensor of Na_vAb(R2G). Grey dashed lines show interatomic distances shorter than 4 Å. See Extended Data Table 1 for details.

blocks the gating pore (Fig. 3f). These structural models illustrate how R2 charge mutations that cause hypokalaemic periodic paralysis result in gating pore current in the resting state.

The Na_vAb(R2S) mutant channel is much more permeant than Na_vAb(R3G) to guanidinium ions[24] (Fig. 1). Guanidinium ions are chemically similar to the distal moiety of the arginine side chain, and guanidine compounds with hydrophobic substituents can block mutant gating pores[24]. We probed our gating pore structures for guanidinium-binding sites by soaking crystals of Na_vAb(R2G) and Na_vAb(R3G) with guanidinium and methylguanidinium to determine whether they would bind in place of the missing side chain of R2 or R3. The crystal structures did not show guanidinium binding to Na_vAb(R3G). However, crystals of Na_vAb(R2G) soaked with guanidinium or methylguanidinium diffracted to 2.7 Å and 2.5 Å resolution, respectively, and unambiguous electron density was observed in place of each R2 side chain (Fig. 3g–i, Extended Data Fig. 5a, b). Bound guanidinium is clearly seen in $2F_o-F_c$ maps (Fig. 3h). E32 and M29 from S1, N49 from S2, R1 and R3 from S4, and Q150 from an adjacent subunit form the binding site for guanidinium (Fig. 3i). M29 and R3 each bind guanidinium through hydrogen bonds (Fig. 3h, i). The carbonyl group of E32 and the carbonyl oxygen of R1 further lock guanidinium in place (Fig. 3h, i). The binding site is flanked by hydrogen bonds from N49 and Q150 that stabilize guanidinium from opposite sides (Fig. 3h, i). The binding site for methylguanidinium is almost identical (Extended

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Data Fig. 5c, d). These structures capture guanidinium bound at a specific site in the closed R2G gating pore. The amino acid residues responsible for guanidinium binding are highly conserved in Na_vAb, $Na_v1.4$ and $Ca_v1.1$ (Extended Data Fig. 1). Substituted guanidinium ions can block gating pore current without major effects on $Na_v1.4$ function[24], suggesting that guanidinium-containing compounds specific for this binding site could provide a basis for structure-based drug design and be used therapeutically to relieve the symptoms of hypokalaemic periodic paralysis.

To examine relationships among structural fluctuations of the gating pore, ionic hydration and Na^+ leakage, we performed molecular dynamics simulations of wild-type and R3G mutant voltage sensors in a hydrated lipid bilayer (Fig. 4). Multiple unbiased simulation repeats, with a total duration of 30 μs, show that the overall structures are conserved. Analysis of axial distributions of water molecules revealed a narrow region ($-5\,Å < z < 5\,Å$) that is more hydrated in the R3G mutant than in the wild-type voltage sensor, owing to the larger size of the lumen in the mutant (Fig. 4a–c, yellow; $P < 0.002$, see Extended Data Table 2). The average count of water molecules within the HCS was 3.9 ± 0.8 and 5.3 ± 0.4 for the wild type and R3G mutant, respectively (Fig. 4e). We performed umbrella-sampling simulations to compute the free energy of Na^+ permeation along the principal axis of the voltage sensor. When Na^+ was within the HCS, the number of water molecules in the HCS increased to 8.4 ± 0.3 in the wild type and 9.0 ± 0.3 in the R3G mutant, respectively. The free-energy profile for Na^+ translocation forms a broad barrier spanning the HCS, centred at Cα of R3. The R3G mutation significantly decreases the height of this barrier from 18 ± 0.8 to 11 ± 1.4 kcal mol^{-1} (Fig. 4e). These values are consistent with the undetectable gating-pore conductance in the wild type and an upper limit of around 0.1 pS in the R3G mutant[29]. Analysis of ionic coordination shows that, at the extracellular edge of the barrier, the first solvation shell of Na^+ is almost exclusively composed of water, consistent with the hydrophobic nature of the bottleneck in the voltage sensor (Fig. 4f). The total coordination number of 5.81 ± 0.02 in bulk water drops to 4.88 ± 0.04 at the peak of the free-energy barrier, suggesting a large desolvation penalty for Na^+ that is partly alleviated by the cavity created in the absence of the R3 side chain. Charge–charge repulsion is also likely to contribute substantially to the higher energy barrier to Na^+ permeation in the wild type, impeding the gating pore leakage observed in the R3G mutant.

The location of R4 coincides with a secondary shoulder in the free-energy profiles (Fig. 4d, R108), indicating that movement of Na^+ past R4 is not rate-limiting for permeation, even though transit of Na^+ past R4 causes the largest displacement of water by protein ligands (Fig. 4e). Spontaneous disruption of the R4–E59 salt bridge in $3 \pm 1\%$ of simulation frames for the wild type and R3G mutant opens the inner end of the gating pore with sufficient frequency to support gating pore current (Fig. 4g–i). Na^+ often makes direct contacts with the anionic side chains of D80 and E59 (Fig. 4f, g), and its movement is coupled to dynamic rearrangements of the R4 salt-bridge network.

Overall, our results provide an unprecedented high-resolution view of functional effects of ion channel mutations that cause periodic paralysis and define the structural basis for pathogenesis of this ion channelopathy. R2G and R3G mutations do not perturb the backbone structure of the voltage sensor, suggesting that the aberrant gating pore currents are not caused by conformational changes in transmembrane alpha helices. Instead, the absence of the positively charged R2 and R3 side chains opens an aqueous gating pore that allows diffusion of Na^+ into the cell, depending on the functional state of the voltage sensor. Our structural studies show how this pathogenic gating pore current is gated in resting and activated states by transmembrane movements of the S4 segment. Although our studies of R2G and R3G mutants suggest a straightforward explanation for the pathogenic gating pore current, mutations that cause hypokalaemic periodic paralysis and normokalaemic periodic paralysis that substitute large side chains such as tryptophan also cause gating pore currents[4,16,17], perhaps by perturbing

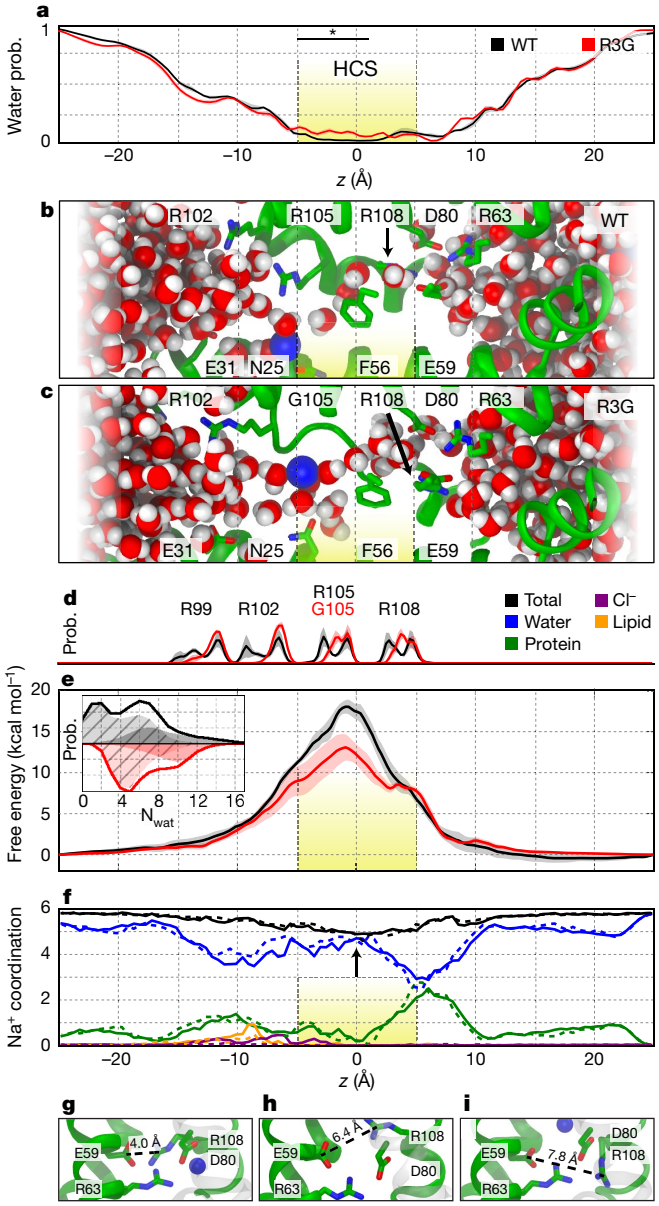


Fig. 4 | R3G mutation lowers the free-energy barrier for Na^+ conductance. a, Probability (prob.) distribution of water along the domain axis for Na_vAb(WT) (black) and Na_vAb(R3G) (red). **b**, **c**, Representation of voltage sensor from Na_vAb(WT) and Na_vAb(R3G) simulations where Na^+ (blue sphere) is restrained at $z = -5$ Å. The S2 segment (residues 45–65) is omitted for clarity. Arrows indicate the positions of R108. **d**, Axial distribution of gating charge Cα for Na_vAb(WT) and Na_vAb(R3G). The axial position in the crystallographic structure is shown as a vertical line. **e**, Inset, probability distribution of water in the HCS (−5 Å to 5 Å) across all simulations of Na_vAb(WT) (black) and Na_vAb(R3G) (red). The total probability is separated into frames where Na^+ occupies the hydrophobic constriction (solid) or is outside this region (cross-hatched). N_{wat}, number of water molecules. Main panel, Potential of mean force for Na^+ conduction within the Na_vAb(WT) (black) and Na_vAb(R3G) (red) pore computed using umbrella sampling. The HCS is highlighted in yellow. **f**, Average coordination of Na^+ as a function of ionic position along the principal axis of the voltage sensor, for Na_vAb(WT) (solid lines) and Na_vAb(R3G) (dashed lines). The first coordination shell of Na^+ is partitioned for coordination to protein (green), water (blue), lipid head groups (orange) and counterions (purple). **g**–**i**, Representative snapshots from simulations of Na_vAb(R3G) depicting conformational isomerization of R4. $^*P < 0.002$, $n = 60$; see Extended Data Table 2 for details. Arrow in **f** indicates the point of minimum coordination of Na^+ by water.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

the local structure of the voltage sensor and thereby opening a pore across the membrane.

Our structures reveal the binding pose of a highly permeant ion, guanidinium, in the closed gating pore of the activated voltage sensor of Na_vAb(R2G). Substituted guanidinium derivatives can block gating pore current without impairing voltage sensor function in $Na_v1.4$[24]. Therefore, our high-resolution structural models may provide molecular templates for design and development of drugs that would mimic guanidinium, block gating pore current and provide symptomatic relief of periodic paralysis.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0120-4.

Received: 18 December 2017; Accepted: 9 April 2018;
Published online: 16 May 2018

1. Venance, S. L. et al. The primary periodic paralyses: diagnosis, pathogenesis and treatment. *Brain* **129**, 8–17 (2006).
2. Cannon, S. C. *Handbook of Experimental Pharmacology* (Springer, Berlin, 2017).
3. Sokolov, S., Scheuer, T. & Catterall, W. A. Gating pore current in an inherited ion channelopathy. *Nature* **446**, 76–78 (2007).
4. Struyk, A. F. & Cannon, S. C. A Na^+ channel mutation linked to hypokalemic periodic paralysis exposes a proton-selective gating pore. *J. Gen. Physiol.* **130**, 11–20 (2007).
5. Vicart, S. et al. New mutations of SCN4A cause a potassium-sensitive normokalemic periodic paralysis. *Neurology* **63**, 2120–2127 (2004).
6. Sokolov, S., Scheuer, T. & Catterall, W. A. Depolarization-activated gating pore current conducted by mutant sodium channels in potassium-sensitive normokalemic periodic paralysis. *Proc. Natl Acad. Sci. USA* **105**, 19980–19985 (2008).
7. Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* **475**, 353–358 (2011).
8. Catterall, W. A. & Zheng, N. Deciphering voltage-gated Na^+ and Ca^{2+} channels by studying prokaryotic ancestors. *Trends Biochem. Sci.* **40**, 526–534 (2015).
9. Catterall, W. A., Goldin, A. L. & Waxman, S. G. International Union of Pharmacology. XLVII. Nomenclature and structure–function relationships of voltage-gated sodium channels. *Pharmacol. Rev.* **57**, 397–409 (2005).
10. Yan, Z. et al. Structure of the $Na_v1.4$-β1 complex from electric eel. *Cell* **170**, 470–482 (2017).
11. Shen, H. et al. Structure of a eukaryotic voltage-gated sodium channel at near-atomic resolution. *Science* **355**, eaal4326 (2017).
12. Catterall, W. A. Ion channel voltage sensors: structure, function, and pathophysiology. *Neuron* **67**, 915–928 (2010).
13. Starace, D. M. & Bezanilla, F. A proton pore in a potassium channel voltage sensor reveals a focused electric field. *Nature* **427**, 548–553 (2004).
14. Sokolov, S., Scheuer, T. & Catterall, W. A. Ion permeation through a voltage-sensitive gating pore in brain sodium channels having voltage sensor mutations. *Neuron* **47**, 183–189 (2005).
15. Gamal El-Din, T. M., Scheuer, T. & Catterall, W. A. Tracking S4 movement by gating pore currents in the bacterial sodium channel NaChBac. *J. Gen. Physiol.* **144**, 147–157 (2014).
16. Jurkat-Rott, K., Groome, J. & Lehmann-Horn, F. Pathophysiological role of omega pore current in channelopathies. *Front. Pharmacol.* **3**, 112 (2012).
17. Moreau, A., Gosselin-Badaroudine, P. & Chahine, M. Biophysics, pathophysiology, and pharmacology of ion channel gating pores. *Front. Pharmacol.* **5**, 53 (2014).
18. Wu, F. et al. A sodium channel knockin mutant ($Na_v1.4$-R669H) mouse model of hypokalemic periodic paralysis. *J. Clin. Invest.* **121**, 4082–4094 (2011).
19. Wu, F. et al. A calcium channel mutant mouse model of hypokalemic periodic paralysis. *J. Clin. Invest.* **122**, 4580–4591 (2012).
20. Gosselin-Badaroudine, P., Delemotte, L., Moreau, A., Klein, M. L. & Chahine, M. Gating pore currents and the resting state of $Na_v1.4$ voltage sensor domains. *Proc. Natl Acad. Sci. USA* **109**, 19250–19255 (2012).
21. Moreau, A., Gosselin-Badaroudine, P., Boutjdir, M. & Chahine, M. Mutations in the voltage sensors of domains I and II of $Na_v1.5$ that are associated with arrhythmias and dilated cardiomyopathy generate gating pore currents. *Front. Pharmacol.* **6**, 301 (2015).
22. Monteleone, S. et al. Mechanisms responsible for ω-pore currents in Ca_v calcium channel voltage-sensing domains. *Biophys. J.* **113**, 1485–1495 (2017).
23. Chakrabarti, N. et al. Catalysis of Na^+ permeation in the bacterial sodium channel Na_vAb. *Proc. Natl Acad. Sci. USA* **110**, 11331–11336 (2013).
24. Sokolov, S., Scheuer, T. & Catterall, W. A. Ion permeation and block of the gating pore in the voltage sensor of $Na_v1.4$ channels with hypokalemic periodic paralysis mutations. *J. Gen. Physiol.* **136**, 225–236 (2010).
25. Payandeh, J., Gamal El-Din, T. M., Scheuer, T., Zheng, N. & Catterall, W. A. Crystal structure of a voltage-gated sodium channel in two potentially inactivated states. *Nature* **486**, 135–139 (2012).
26. Yarov-Yarovoy, V. et al. Structural basis for gating charge movement in the voltage sensor of a sodium channel. *Proc. Natl Acad. Sci. USA* **109**, E93–E102 (2012).
27. Vargas, E. et al. An emerging consensus on voltage-dependent gating from computational modeling and molecular dynamics simulations. *J. Gen. Physiol.* **140**, 587–594 (2012).
28. Catterall, W. A., Wisedchaisri, G. & Zheng, N. The chemical basis for electrical signaling. *Nat. Chem. Biol.* **13**, 455–463 (2017).
29. Cooper, K. E., Gates, P. Y. & Eisenberg, R. S. Diffusion theory and discrete rate constants in ion permeation. *J. Membr. Biol.* **106**, 95–105 (1988).

Acknowledgements We thank the beamline staff at the Advanced Light Source (BL8.2.1 and BL8.2.2) for assistance during data collection and J. Li for technical and administrative assistance. This research was supported by National Institutes of Health research grants R01 NS015751 (W.A.C.) and R01 HL112808 (W.A.C. and N.Z.), by the Howard Hughes Medical Institute (N.Z.), and by Canadian Institutes of Health Research grant MOP 130461 (R.P.).

Author contributions D.J., T.M.G.E.-D., C.I., P.L., R.P., N.Z. and W.A.C. designed experiments. D.J., T.M.G.E.-D., C.I. and P.L. conducted experiments. D.J., T.M.G.E.-D., C.I., R.P., N.Z. and W.A.C. analysed the results. T.M.G.E.-D., C.I. and W.A.C. wrote the paper with input from all co-authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0120-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0120-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to N.Z. or W.A.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Electrophysiology. All experiments were performed using *T. ni* insect cells (High Five Cells, Thermofisher). Molecular biology and patch-clamp measurements were performed as described previously[15,30]. All constructs showed high level expression that enabled us to measure ionic current and gating pore currents 48 h after infection. Whole-cell sodium currents were recorded using an amplifier (Axopatch 200; Molecular Devices) with glass micropipettes (2–4 MΩ). The intracellular pipette solution contained (mM): 35 NaCl, 105 CsF, 10 EGTA and 10 HEPES, pH 7.4 (adjusted with CsOH). The extracellular solution contained (mM): 140 NaCl, 2 $CaCl_2$, 1.8 $MgCl_2$ and 10 HEPES, pH 7.4 (adjusted with NaOH).

For Na_vAb(R2S), the standard clamp protocol for measuring central pore currents consisted of steps from a holding potential of −200 mV to voltages ranging from −180 to 0 mV in 10 mV steps. For Na_vAb(R3G), cells were held at −160 mV and 10 mV voltage steps ranging from −140 mV to +50 mV were applied. A P/−10 or P/−4 leak-subtraction protocol was used to subtract linear leak and capacitive currents from holding potentials of −200 or −160 mV, respectively.

To measure gating pore currents in Na_vAb(R2S), cells were held at −200 mV for ∼1 min to allow recovery from slow inactivation. Then, the cells were held at −100 mV for gating pore current measurements, which inactivates the central pore current. Depolarizing pulses in 10-mV steps were applied from −200 mV up to +50 mV. The intracellular pipette solution contained (mM): 140 CsF, 10 EGTA, and 10 HEPES, pH 7.4 (adjusted with CsOH). The extracellular solution contained (mM): 140 NaCl, 2 $CaCl_2$, 1.8 $MgCl_2$ and 10 HEPES, pH 7.4 (adjusted with NaOH). To test gating pore selectivity for different cations, NaCl was replaced by an equimolar concentration of KCl, CsCl, LiCl, NMDG or 40 mM guanidinium sulfate +100 mM NMDG, 40 mM methylguanidinium sulfate +100 mM NMDG, or 40 mM ethylguanidinium sulfate +100 mM NMDG.

To measure gating pore currents in Na_vAb(R3G), cells were held at 0 mV for a few min to induce slow inactivation. Then, 10-mV pulses were applied from −200 mV up to +50 mV. To measure ion selectivity of R3G, the composition of external solution was in (mM): 140 NMDG-MS, 2 $CaCl_2$, 10 HEPES. The intracellular solution contains either 140 mM NaF, 140 mM KF or 140 mM CsF, in addition to 10 mM HEPES, 10 mM EGTA.

No online leak subtraction protocols were used during measuring of gating pore currents. Linear leak subtraction was done offline by generating a linear fit to the *I–V* curves at voltage ranges +100 mV to 0 mV for Na_vAb(R2S) and between −200 mV and 0 mV Na_vAb(R3G). Voltage-clamp pulses were generated and currents were recorded using Pulse software controlling an Instrutech ITC18 interface (HEKA). Data were analysed using Igor Pro 6.37 software (WaveMetrics).

Protein purification and crystallization. R2G, R2S or R3G mutations were introduced into Na_vAb/I217C by site-directed mutagenesis (QuikChange; Agilent) and confirmed by sequencing. Protein was expressed and purified as described[6]. In brief, recombinant baculovirus was generated by using the Bac-to-Bac system (Invitrogen), and *T. ni* cells (High Five Cells, Thermofisher) were infected for protein production. Protein was extracted with 1% digitonin (EMD Biosciences). After centrifugation, the supernatant was agitated with anti-Flag M2–agarose resin (Sigma). Flag resin was washed and eluted with Flag peptide, and the purified protein was analysed by SDS–PAGE (Extended Data Fig. 6). Purified protein was then loaded onto a Superdex 200 column (GE Healthcare) in 10 mM Tris-HCl pH 8.0, 100 mM NaCl and 0.12% digitonin. The peak fraction was concentrated to ∼17 mg ml^{-1} and reconstituted into DMPC:CHAPSO (Anatrace) bicelles. The protein–bicelle preparation was mixed in a 1:1 ratio and set in a hanging-drop vapour-diffusion format over a well solution containing 1.8–2.0 M ammonium sulfate, 100 mM Na-citrate pH 4.8–5.2. Crystals grew to full size in a week. Crystals were cryoprotected in well solution supplemented with 28% glucose (w/v) in increments of 7% glucose during harvesting. Guanidinium- or methylguanidinium-bound crystals were cryoprotected by soaking in the same cryoprotection solution plus 10 mM guanidinium or methylguanidinium ions. Crystals were plunged into liquid nitrogen for data collection.

Data collection and structure determination. X-ray diffraction data was collected at Advanced Light Source (beamlines BL821 and BL822), and then integrated and scaled with the HKL2000 suite. Both Na_vAb(R2G) and Na_vAb(R3G) structures were solved by Phaser-MR using Na_vAb (PDB code: 3RVY) monomer as searching model. After initial phases, models were refined with PHENIX[31] and manually re-built using COOT[32]. High-resolution density maps clearly showed no side-chain density for R2G or R3G. Simulated annealing omit maps were used to confirm the binding of guanidinium ions. The geometries of the final models were verified using MolProbity[33]. All solvent-accessible volume analysis in the voltage-sensing modules was generated with MOLE2[34].

Molecular modelling and dynamics. Molecular models of the Na_vAb(WT) and Na_vAb(R3G) channels were constructed using the Na_vAb(I217C) structure (PDB code: 3RVY)[6]. The latter model was generated by substituting R105 with G in all four voltage sensing domains. Both systems were embedded in a hydrated 1,2-dimyristoyl-*sn*-glycero-3-phosphatidylcholine (DMPC) bilayer with ∼250 mM NaCl for a total of ∼129,000 atoms. Embedding was performed using the alchembed protocol[35] using an equilibrated rectangular CHARMM36 DMPC bilayer patch obtained from the Klauda laboratory website (https://terpconnect.umd.edu/∼jbklauda/). The protein, lipids and ions were modelled with the CHARMM36 all-atom force field[36–38] and water molecules were modelled with TIP3P[39]. NBFIX adjustments were made for Na^+-backbone carbonyl O atom and Na^+-lipid head group interactions[40,41].

All simulations were performed with GROMACS 5.0.6[42]. Electrostatic interactions were calculated using particle-mesh Ewald algorithm[43,44] with a real-space cut-off distance of 1.2 nm, a grid spacing of 0.16 nm and cubic interpolation. Lennard–Jones interactions were cut off at 1.2 nm. Nonbonded interactions were calculated using Verlet neighbour lists[45]. All simulations were performed at constant temperature (300 K) using the Nosé–Hoover thermostat[46,47] with temperature coupling of 0.5 ps and at constant pressure (1 atm) with the Parrinello–Rahman barostat[48,49] with a time constant of 2 ps. All chemical bonds were constrained using the LINCS algorithm[50]. The integration timestep was 2 fs.

Because the channel and voltage sensor were initially devoid of water molecules and ions, a protein-restrained equilibration period of 30 ns was used to reduce the systematic sampling bias induced by the initial conditions (10 ns with protein heavy-atom restraints, 10 ns with backbone restraints, and 10 ns with Cα restraints, all with a force constant of 2.39 kcal mol^{-1} $Å^{-2}$). Unbiased production simulations of 15 replicas of 'WT' and 'R3G' systems were conducted for 1,000 ns each, resulting in aggregate sampling of 15 μs for each tetramer (4 × 15 μs = 60 μs for WT and R3G voltage sensors).

Simulation snapshots beyond $t = 100$ ns were extracted from unbiased simulations and used as initial conditions for biased simulations, using the entire tetramer. Umbrella sampling[51,52] was used to compute the free energy or potential of mean force (PMF) profile for the movement of Na^+ through voltage sensing domain. The range of the reaction coordinate, −2.0 to 2.0 nm with respect to the centre of the hydrophobic constriction, was discretized into ∼130 unevenly spaced windows. For each window, biased simulations were initiated with a water molecule exchanged for Na^+ in all four voltage sensors. Production simulations were performed for 70–100 ns per window with a harmonic restraining potential force constant of 2.39 kcal mol^{-1} $Å^{-2}$ and a flat-bottom cylindrical position restraint for all four Na^+ ions simultaneously. The axial position of the permeating Na^+ ion, z, was stored every 10 fs and the data from each of the four voltage sensors were used separately to generate four independent PMF profiles using g_wham[53], enforcing cyclic periodicity of the PMF in the bulk (at $z = -2.5$ nm). The initial 10 ns were excluded from each umbrella sampling run. We report the mean PMF over the four voltage sensors with error bars computed using the standard error of mean over all four PMFs. The total simulation time for each of the two systems (WT and R3G) was ∼11 μs, yielding a total of ∼45 μs of voltage sensor data.

Water occupancy of the voltage sensor was computed by counting the number of water oxygen atoms within a cylinder of radius 8.0 Å. We define the hydrophobic constriction centre as the geometric centre of Cα atoms of residues 22, 57, 84 and 105. The range of the HCS is defined as −5 Å to 5 Å along the axial coordinate of the voltage sensor. Coordination of Na^+ to channel ligands, water, ions and lipids was performed by computing the number of protein, water and lipid O atoms, as well as Cl^- ions, within the first solvation shell of Na^+ (<3.0 Å). The average coordination number at a given axial position was computed over all simulation frames regardless of the subunit, but the total coordination number in bulk water and at the hydrophobic constriction reported in the text was based on the mean and standard error of mean over the four voltage sensors. Analysis of the trajectories was performed using MDTraj[54] and molecular renderings were generated using Visual Molecular Dynamics[55].

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

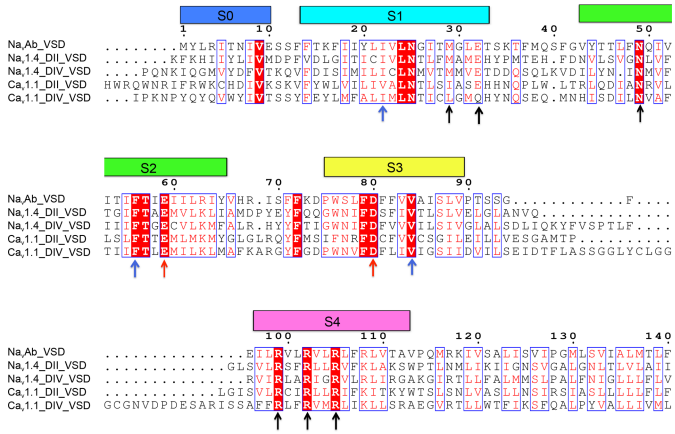
Data availability. Coordinates and structure factors have been deposited in the Protein Data Bank with the following accession numbers: Na_vAb(R3G), 6C1E; Na_vAb(R2G)–guanidinium, 6C1K; Na_vAb(R2G)–methylguanidinium, 6C1M; Na_vAb(R2G), 6C1P.

30. Gamal El-Din, T. M., Martinez, G. Q., Payandeh, J., Scheuer, T. & Catterall, W. A. A gating charge interaction required for late slow inactivation of the bacterial sodium channel Na_vAb. *J. Gen. Physiol.* **142**, 181–190 (2013).
31. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
32. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
33. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
34. Petřek, M., Kosinová, P., Koca, J. & Otyepka, M. MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure* **15**, 1357–1363 (2007).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

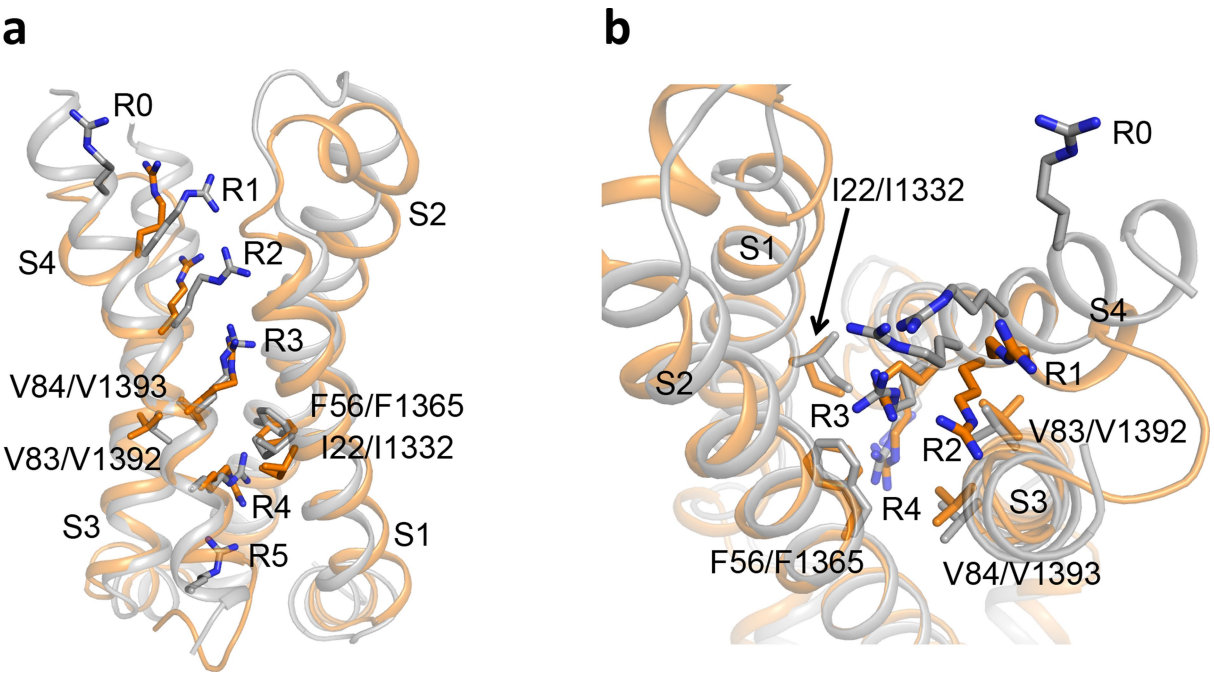
35. Jefferys, E., Sands, Z. A., Shi, J., Sansom, M. S. P. & Fowler, P. W. Alchembed: a computational method for incorporating multiple proteins into complex lipid geometries. *J. Chem. Theory Comput.* **11**, 2743–2754 (2015).
36. MacKerell, A. D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
37. Klauda, J. B. et al. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* **114**, 7830–7843 (2010).
38. Best, R. B. et al. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ_1and χ_2 dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).
39. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
40. Noskov, S. Y. & Roux, B. Control of ion selectivity in LeuT: two Na^+ binding sites with two different mechanisms. *J. Mol. Biol.* **377**, 804–818 (2008).
41. Venable, R. M., Luo, Y., Gawrisch, K., Roux, B. & Pastor, R. W. Simulations of anionic lipid membranes: development of interaction-specific ion parameters and validation using NMR data. *J. Phys. Chem. B* **117**, 10183–10192 (2013).
42. Abraham, M. J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *Software* ***X*** **1–2**, 19–25 (2015).
43. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an *N*·log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
44. Essmann, U. et al. A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).
45. Verlet, L. Computer 'experiments' on classical fluids. I. thermodynamical properties of Lennard–Jones molecules. *Phys. Rev.* **159**, 98–103 (1967).
46. Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* **31**, 1695–1697 (1985).
47. Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **52**, 255–268 (1984).
48. Parrinello, M. & Rahman, A. Crystal structure and pair potentials: A molecular-dynamics study. *Phys. Rev. Lett.* **45**, 1196–1199 (1980).
49. Nosé, S. & Klein, M. L. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.* **50**, 1055–1076 (1983).
50. Hess, B. P-LINCS: a parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* **4**, 116–122 (2008).
51. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J. Comput. Phys.* **23**, 187–199 (1977).
52. Roux, B. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* **91**, 275–282 (1995).
53. Hub, J. S., de Groot, B. L. & van der Spoel, D. g_wham—a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *J. Chem. Theory Comput.* **6**, 3713–3720 (2010).
54. McGibbon, R. T. et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
55. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996). 27–28.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



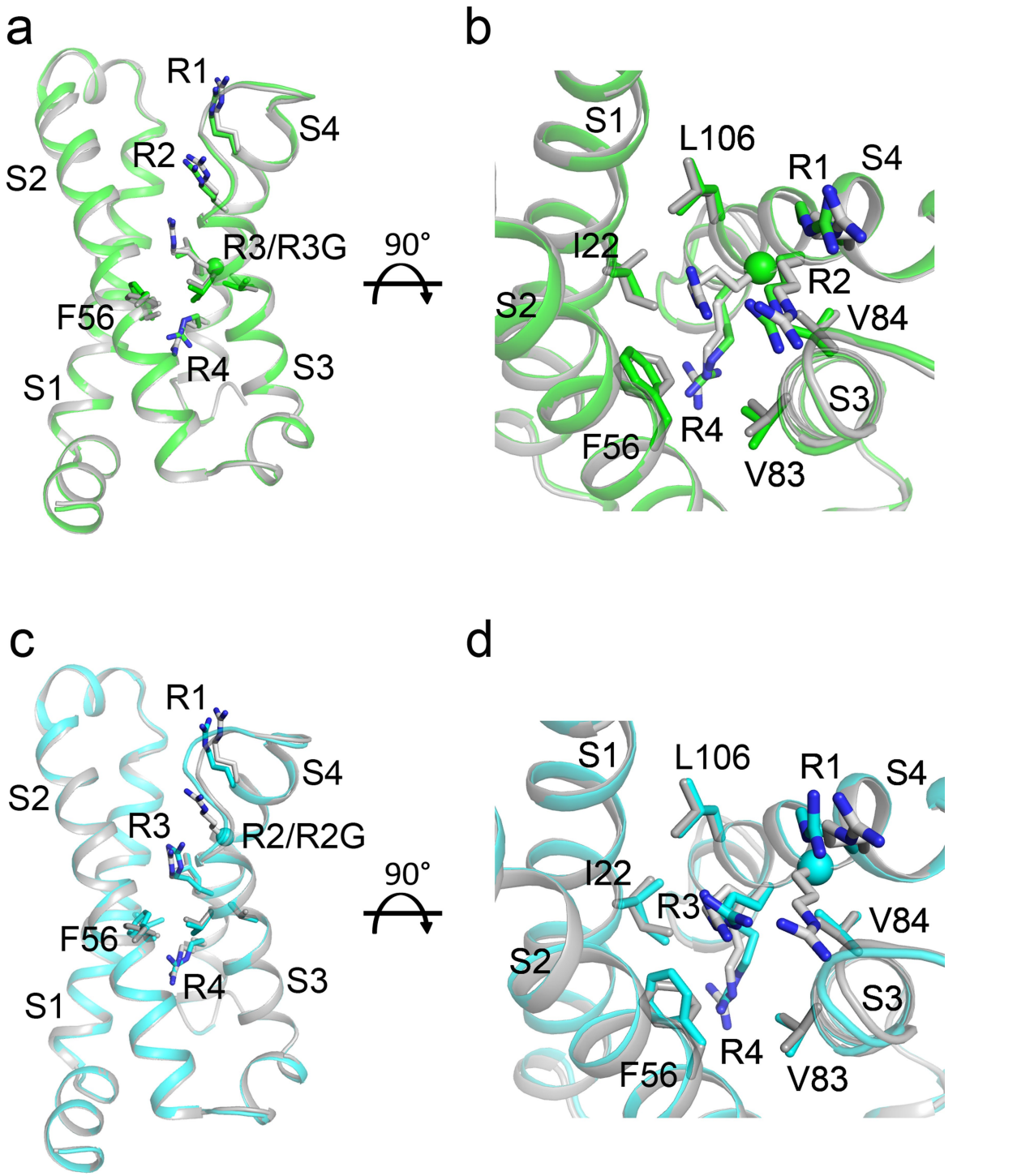
Extended Data Fig. 1 | Sequence alignment of the voltage sensor of Na_vAb with those of human $Na_v1.4$ homologous domain (D)II, $Na_v1.4$ DIV, $Ca_v1.1$ DII and $Ca_v1.1$ DIV. Coloured rectangles represent transmembrane helices. Black arrows indicate residues that form the guanidinium binding site, blue arrows indicate the hydrophobic constriction site and red arrows indicate the conserved intracellular negative cluster.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



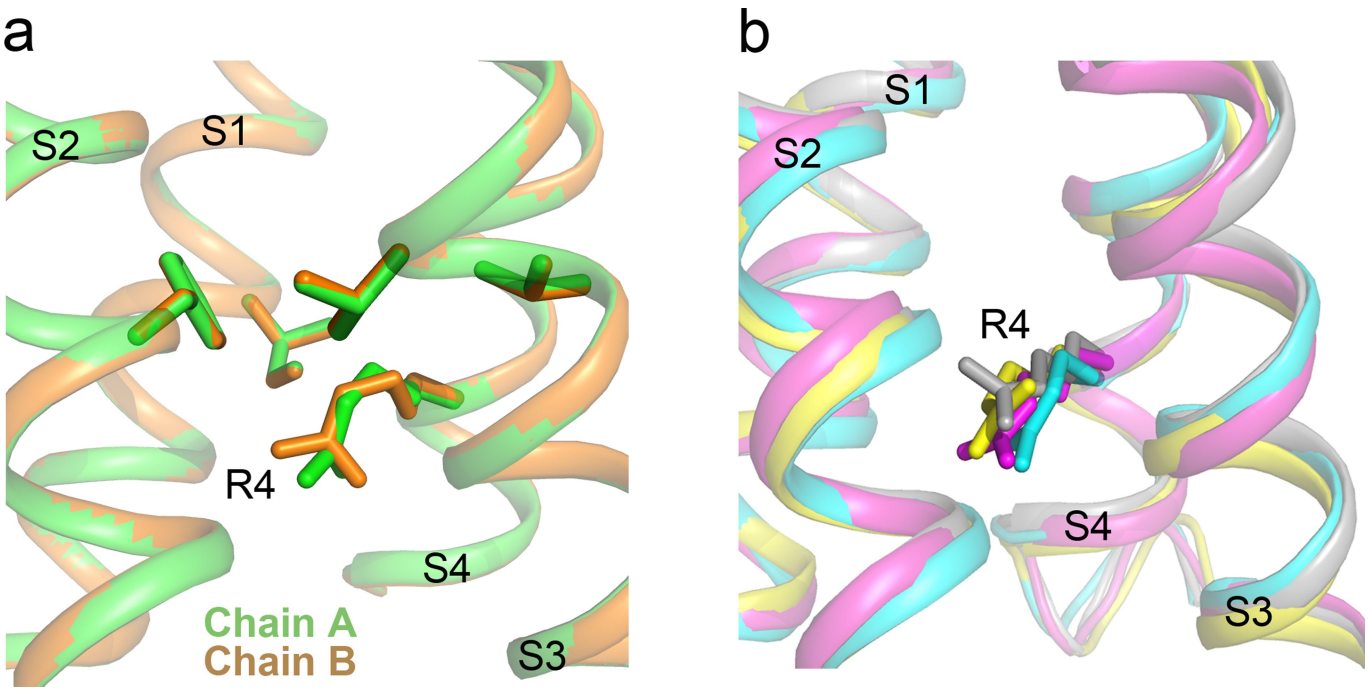
Extended Data Fig. 2 | Superposition of the Na_vAb(WT) voltage sensor and the *Electrophorus electricus* (electric eel) $Na_v1.4$ DIV voltage sensor. **a**–**b**, Comparison of the conformations of Na_vAb(WT) voltage sensor (orange) and *Ee*$Na_v1.4$ voltage sensor DIV (PDB code: 5XSY) (grey) in side view and top view, respectively. Arginine sensors and hydrophobic residues in the HCS are labelled and shown with side chains in sticks.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



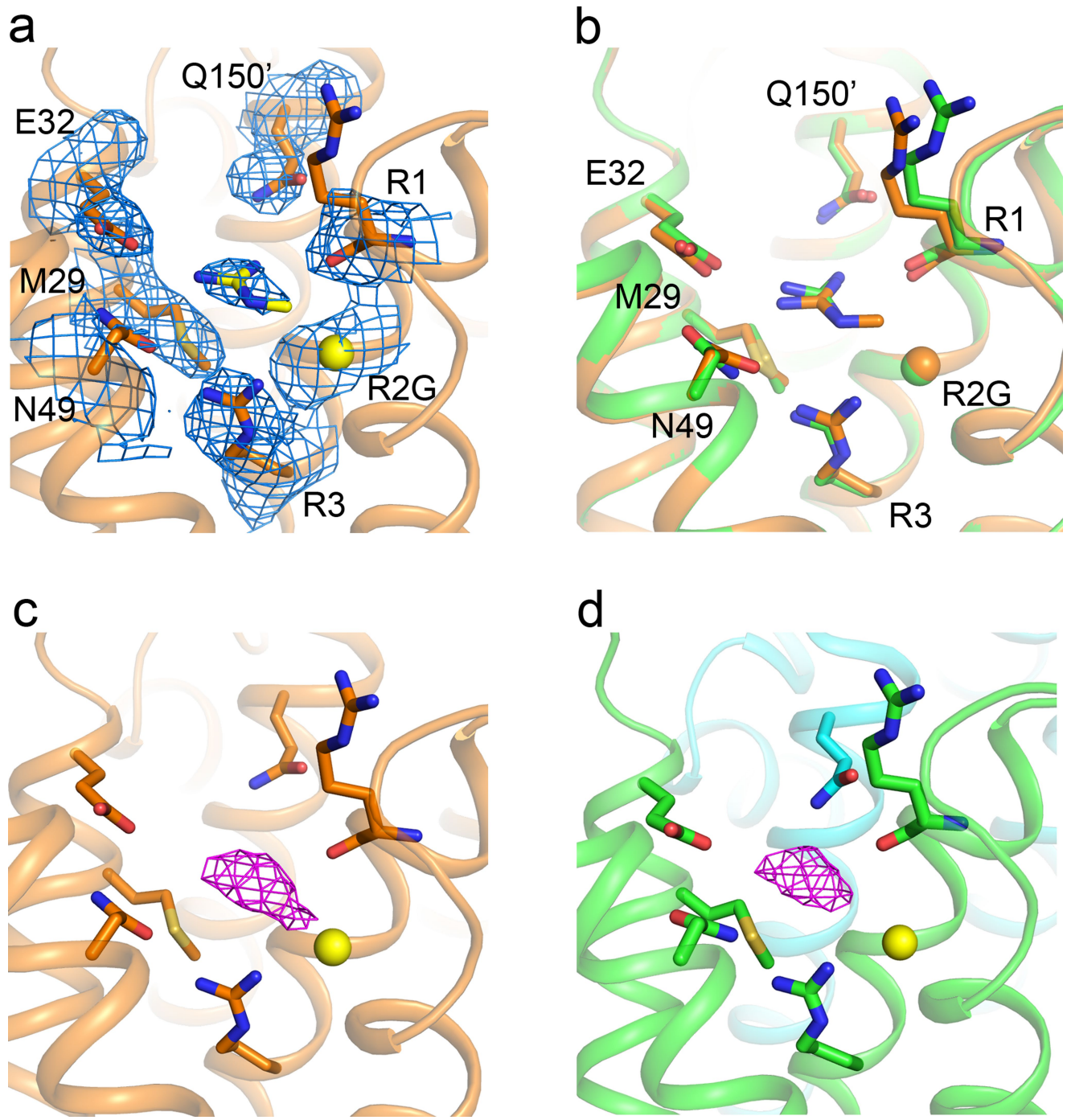
Extended Data Fig. 3 | Superposition of the voltage sensors of Na_vAb(WT) and mutant channels. a–b, Voltage sensor structure alignment between Na_vAb(WT) (grey) and Na_vAb(R3G) (green) in side view and top view, respectively. **c–d**, Voltage sensor structure alignment between Na_vAb(WT) (grey) and Na_vAb(R2G) (cyan) in side view and top view, respectively. Arginine sensors and hydrophobic residues in the HCS are labelled and shown with side chains in sticks.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



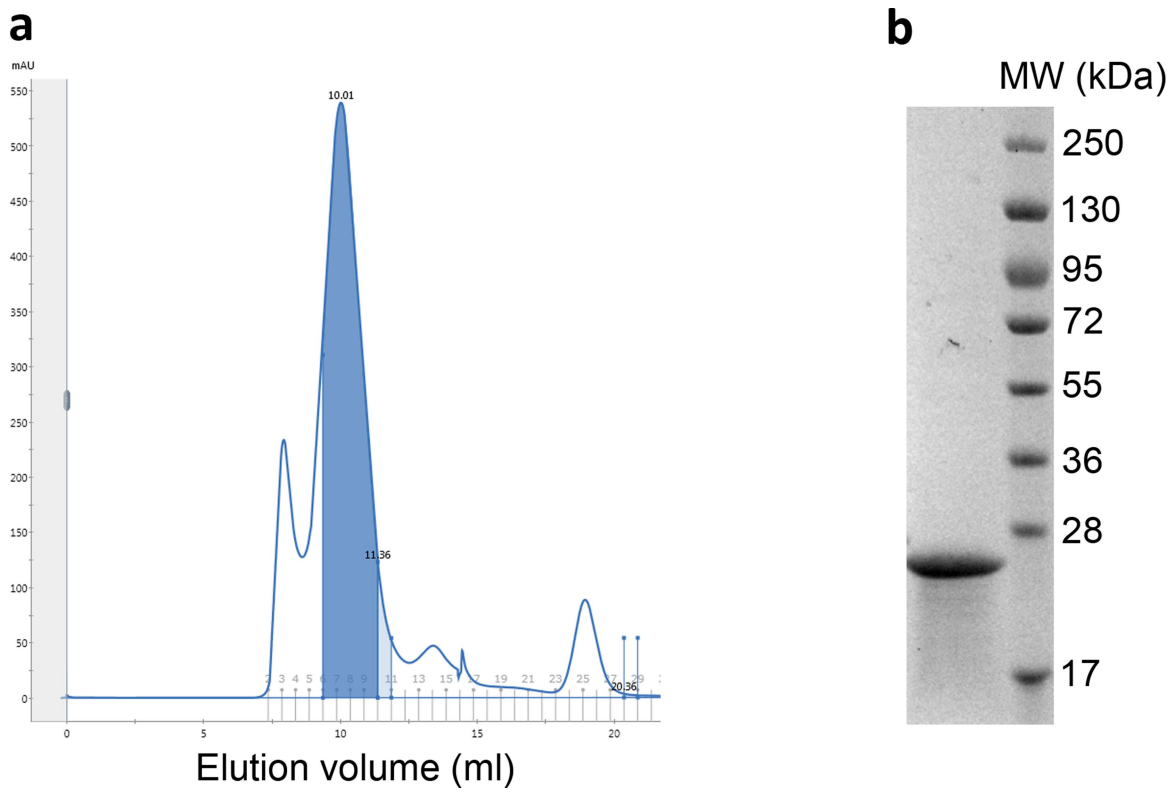
Extended Data Fig. 4 | R4 side chain conformational changes. **a**, Different conformations of the R4 rotamer in Na_vAb(R3G) chain A (green) and chain B (orange). **b**, Different conformations of the R4 rotamer in the four subunits of Na_vAb in the slow-inactivated state (PDB code: 4EKW).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 5 | Electron density maps for bound guanidinium and methylguanidinium ions. a, 2mF_o−DF_c electron density map (blue mesh) of residues around the methylguanidinium binding site at 1σ. **b**, Overlay of guanidinium binding site (green) and methylguanidinium binding site (orange). **c**–**d**, Simulated annealing map (F_o−F_c) contoured at 3σ for methylguanidinium and guanidinium, respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Purification of $Na_vAb(R3G)$. a, Representative gel-filtration chromatography of $Na_vAb(R3G)$; highlighted peak fractions were concentrated for crystallization. **b**, Concentrated sample was visualized on SDS–PAGE by Coomassie blue staining.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Data collection and refinement statistics

	Na_vAb/R3G	Na_vAb/R2G Guanidinium	Na_vAb/R2G Methyl Guanidinium	Na_vAb/R2G Apo
Data collection				
Space group	*I222*	*I222*	*I222*	$P2_122_1$
Cell dimensions				
a, *b*, *c* (Å)	126.8, 127.0, 192.3	126.6, 126.6, 191.8	126.3, 126.2, 191.6	125.5, 125.6, 192.0
α, *β*, *γ* (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90
Wavelength (Å)	0.99994	0.99994	0.99994	0.99994
Resolution (Å)	50-2.90 (3.00-2.90)	50-2.70 (2.80-2.70)	50-2.55 (2.64-2.55)	50-2.80 (2.90-2.80)
Rpim	4.6 (62.6)	4.0 (62.0)	3.9 (64.0)	5.3 (58.1)
$I/\sigma I$	16.6 (1.5)	18.5 (1.2)	18.5 (1.0)	14.5 (0.8)
Completeness (%)	100 (99.9)	99.6 (96.5)	99.4 (95.0)	98.0 (81.6)
Redundancy	7.3 (7.2)	7.1 (5.4)	5.3 (3.8)	5.1 (3.2)
Refinement				
Resolution (Å)	42.50-2.86	42.31-2.70	42.31-2.52	48.46-2.90
No. reflections	35059	41173	51039	67766
R_{work}/ R_{free}	21.25/23.99	20.98/24.59	20.31/22.66	23.35/26.03
No. atoms				
Protein	3606	3605	3673	7160
Ligand/ion	512	449	660	415
Water	0	5	35	0
B-factors				
Protein	108.7	97.8	103.1	112.89
Ligand/ion	128.2	107.5	130.8	115.8
Water		54.5	75.5	
R.m.s deviations				
Bond lengths (Å)	0.010	0.010	0.009	0.012
Bond angles (°)	1.311	1.215	1.253	1.703
Ramachandran plots				
Favored	93.2%	92.5%	94.0%	92.1%
Allowed	6.8%	7.3%	5.4%	7.4%
Outliers	0.0%	0.2%	0.6%	0.5%

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Statistical analysis of voltage sensor water occupancy from molecular simulations

Degrees of freedom	Axial Interval (Å,Å)	t-statistic	q value
60+60-2	(-20,-19)	1.411	2.221E-01
	(-19,-18)	2.878	1.024E-02
	(-18,-17)	4.389	7.674E-05
	(-17,-16)	4.802	1.696E-05
	(-16,-15)	4.545	4.465E-05
	(-15,-14)	4.249	1.148E-04
	(-14,-13)	2.181	5.422E-02
	(-13,-12)	0.740	5.420E-01
	(-12,-11)	0.217	8.721E-01
	(-11,-10)	2.760	1.276E-02
	(-10,-9)	4.283	1.078E-04
	(-9,-8)	-0.110	9.364E-01
	(-8,-7)	1.914	9.279E-02
	(-7,-6)	-5.668	4.626E-07
	(-6,-5)	-5.839	2.674E-07
	(-5,-4)	**-9.376**	**6.032E-15**
	(-4,-3)	**-12.075**	**9.500E-21**
	(-3,-2)	**-11.945**	**9.674E-21**
	(-2,-1)	**-10.018**	**2.422E-16**
	(-1,0)	**-5.812**	**2.674E-07**
	(0,1)	**-7.910**	**1.027E-11**
	(1,2)	1.488	1.993E-01
	(2,3)	1.813	1.073E-01
	(3,4)	2.797	1.205E-02
	(4,5)	5.497	9.078E-07
	(5,6)	2.476	2.672E-02
	(6,7)	-3.688	8.593E-04
	(7,8)	-8.074	5.198E-12
	(8,9)	-3.257	3.462E-03
	(9,10)	-1.302	2.522E-01
	(10,11)	1.220	2.809E-01
	(11,12)	0.482	7.211E-01
	(12,13)	-2.870	1.024E-02
	(13,14)	1.919	9.279E-02
	(14,15)	-0.003	9.978E-01
	(15,16)	0.887	4.569E-01
	(16,17)	1.843	1.044E-01
	(17,18)	1.319	2.522E-01
	(18,19)	0.263	8.686E-01
	(19,20)	0.249	8.686E-01

Results of two-way *t*-tests on differences in average water count in 1 Å segments of the voltage sensor axial coordinate comparing wild-type and R3G mutant simulations. In each segment, we compare the mean of 60 values ($n = 60$, obtained from pooling the mean water counts of the four voltage sensor proteins from each of 15 simulation repeats). The HCS region (−5 to 1 Å, bold) has the largest effect size, indicating a region of biological significance.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

TECHNOLOGY FEATURE

SINGLE-CELL APPROACHES TO IMMUNE PROFILING

Protein- and sequencing-based technologies are helping researchers to profile immune cells ever more deeply.



THOMAS DEERINCK, NCMIR/SPL

Analysing individual white blood cells can help scientists to identify their specific role in fighting disease.

BY ESTHER LANDHUIS

The human immune system is a vast, decentralized army. With billions of specialized cells in constant motion, it's incredibly difficult to work out which ones are doing what in any given place or situation. But advances in proteomics and genomics technologies are helping researchers to catalogue the various parts. It is a massive undertaking that is likely to produce 100 billion times more data than the Human Genome Project.

For decades, immunologists have relied heavily on flow cytometry, a technique that involves labelling cells with different fluorescent markers. Over the past eight to nine years, mass cytometry, which uses metal ions to label the cells, has offered more-detailed glimpses into immune responses to diseases such as cancer, tuberculosis and malaria.

Yet these techniques barely scratch the surface in determining the precise work of T cells and B cells. These cells are adorned with unique protein molecules — known as T-cell receptors and B-cell receptors — that are exquisitely designed to "recognize specific targets and respond to them in an evolutionary way", says Jennifer Sims, a molecular biologist at the Memorial Sloan Kettering Cancer Center in New York City. Immune cells are almost as varied as people are, she says.

These surface receptors recognize specific molecular features, or antigens, on pathogenic organisms or tumours. When a receptor detects a harmful antigen, it triggers the immune cell to multiply and mobilize for attack. Every receptor is distinct, encoded by combinations of gene segments that are shuffled and recombined as the immune cell develops. Last month, at the annual meeting of the American Association for Cancer Research in Chicago, Illinois, Sims chaired a session on methods for analysing the T-cell repertoire — a set of disease-fighting cells that churn out up to 10^{20} unique receptors over a lifetime (see 'A focus on immune receptors').

By studying these receptors — and their B-cell analogues — researchers hope to learn how the immune system responds and evolves during disease. Whereas flow cytometry can reveal the coarse details of how ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

different cell populations rise and fall, the immune-receptor profiling methods can pinpoint the B-cell and T-cell 'clones' involved. To study these specialized receptors and further expand the power of single-cell analyses, a wave of sequencing-based methods is rippling through the field, says Alex Shalek, a chemical physicist at the Massachusetts Institute of Technology (MIT) in Cambridge, whose lab builds tools to study how cells interact in healthy and diseased states. Just as DNA's four 'letters' can encode any gene in the human body, using unique DNA snippets to tag antibodies paves the way for experiments surveying a theoretically infinite number of proteins, Shalek says.

Plus, such technologies enrich immune analyses by allowing researchers to also measure the transcriptome — the complete set of expressed genes — in the same cells. Although it is less-established than mass cytometry, sequencing-based immune profiling can be done in any lab, hospital or clinic with a DNA sequencer, potentially expanding the technology's reach. And researchers can now choose from at least half a dozen commercial systems and do-it-yourself methods for converting RNA and proteins from individual cells into strings of letters that can be read on a standard DNA sequencer.

Collectively, these new technologies are fine-tuning scientists' understanding of how the various different immune cells act in the microenvironment during disease states, says Sims.

FLUORESCENCE TO METALS

Flow cytometry is an optical technique that classifies cells on the basis of properties such as size, granularity and the presence of signature proteins labelled with fluorescent antibodies. 'Flow' refers to the mechanics of the technique: cells flow in single-file past a series of lasers and detectors, which read them as they pass.

The technique has been a staple of immunology for decades, but it has shortcomings. The visible-light spectrum limits most experiments to no more than a dozen or so protein markers — too few for analyses that involve small numbers of cells or complex phenotypes.

In 2009, chemist Scott Tanner, then at the University of Toronto in Canada, came up with a solution[1]: mass cytometry (often called CyTOF) blends flow cytometry with mass spectrometry, using metal-conjugated antibodies to boost the number of detectable markers to 50 or so.

The technique gained popularity after geneticist Garry Nolan and his colleagues at California's Stanford University used it to measure 34 parameters at once in human bone-marrow cells — helping them to track simultaneously how a wide variety of immune cells were responding to different drugs[2]. It "lets us dive into really complicated systems with single-cell heterogeneity and rare [cell] populations, and I don't need to know ahead of time where the action is happening", says Sean Bendall, a Stanford biochemist who worked with Nolan on the profiling. CyTOF was originally commercialized by DVS Sciences, which was acquired by Fluidigm in South San Francisco, California, in 2014.

"It's a very fast-evolving field with a lot of innovation."

Researchers doing flow cytometry will typically choose a small set of protein markers to stain each kind of immune cell — for example, they may use CD19 to identify B cells, and CD4 or CD8 for T-cell subsets. However, this provides only a coarse level of resolution because the excitation and emission spectra of the fluorescent dyes overlap and muddy the picture. So when Asya Rolls, a neuroimmunologist at the Technion Israel Institute of Technology in Haifa, wanted a comprehensive look at all the immune-cell populations in the mouse brain, she decided to use CyTOF instead[3].

She surveyed 44 antibodies simultaneously, and the approach yielded a surprise: a T-cell subpopulation with unusually high expression of the surface protein CD86 (ref. 4). CD86 isn't usually found on T cells in the rest of the body — it typically decorates the surface of other immune cells and acts to regulate T-cell activation — so the build-up in the brain was intriguing. Mass cytometry is a good tool for discovery, Rolls says.

Once mass cytometry reveals which cells and molecules to focus on, however, a flow cytometer often proves more useful for follow-up analysis. Flow cytometry is faster: it can sort more than 10,000 cells per second, whereas mass cytometry handles only about 1,000. CyTOF experiments also tend to use many more antibodies per sample, so any mistakes are costlier. And if you will still need the cells after the analysis, CyTOF is not an option because it vaporizes them before reading their metal tags.

But in its favour, mass spectrometry can target many more parameters per cell — including chemical compounds, collectively called the epigenome, that attach to DNA and influence how a cell uses its genetic instructions. Targeting the epigenome, a Stanford team found last month that immune cells from older people have much greater epigenetic heterogeneity than do those of younger adults[5]. These data support the long-held assumption that the human immune system deteriorates with age because of cell-to-cell variations in gene expression.

DNA-BASED PROFILING

The initial wave of mass-cytometry papers in the early part of this decade caught the attention of Adam Abate. Abate is a physicist at the University of California, San Francisco, who builds tools known as microfluidics devices, which are used to miniaturize the work and form a parallel with biological experiments. "I was reading Garry Nolan's papers," he says. "We had journal clubs on them." Those studies convinced Abate of the value of simultaneously analysing many parameters in the same cell, a process known as multiplexing. The advent of mass cytometry had already boosted the number of protein markers it was possible to analyse in one shot, from 10–15 to perhaps as many as 100. But Abate started asking, "why stop at a hundred? Why not go bigger?" Human cells have around 20,000 protein-coding genes, and more than 100,000 protein variants. "We need more multiplexing," he decided.

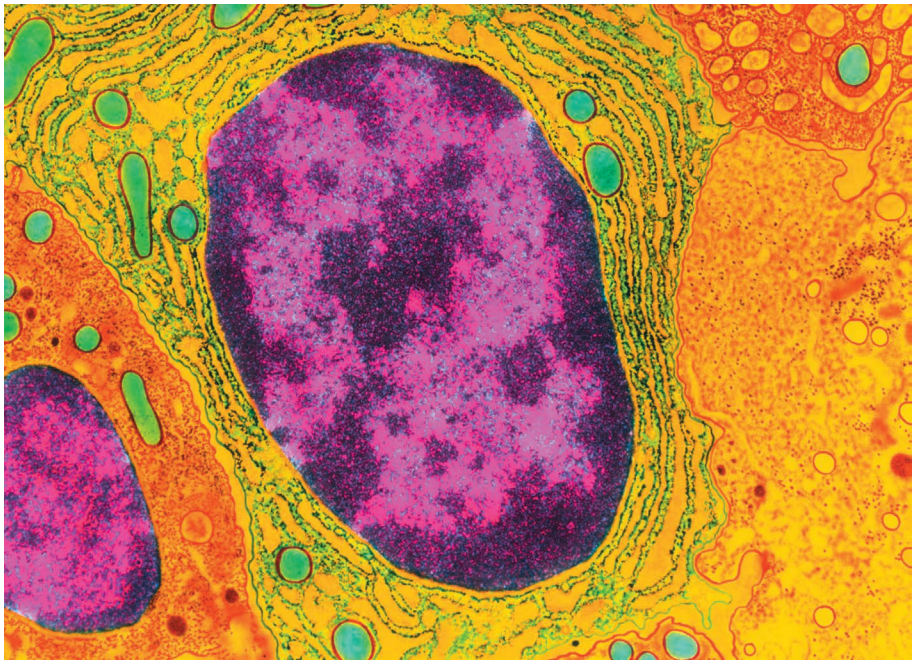
A focus on immune receptors

Trying to pinpoint which of the trillions of T-cell and B-cell receptors in the immune repertoire contribute to a person's disease might seem like an impossible task. But specialized sequencing approaches are enabling researchers to give it a try. About half a dozen companies now offer services or kits for reading out the shuffled gene sequences of T-cell and B-cell receptors from cell samples. By tallying those sequences and watching how they change, researchers can track which clones of cells have key roles in disease.

Adaptive Biotechnologies in Seattle, Washington, offers the whole process. Researchers can send in chunks of frozen tissue and pay for the DNA isolation. They can also submit purified DNA or RNA samples and have the company run the analyses and send back processed data — spreadsheets with lists of nucleotide sequences identified in that run, and their frequencies. Other companies support a more do-it-yourself approach. ArcherDx in Boulder, Colorado; Takara Bio in Mountain View, California; iRepertoire in Huntsville, Alabama; and New England Biolabs in Ipswich, Massachusetts, sell multiplex PCR primers that researchers can use to amplify T- and B-cell receptor gene segments from their own samples.

Each system has pros and cons. Some are great for frozen tissue, others for formalin-fixed, paraffin-embedded samples. Still others are sensitive to interference from tumour-cell DNA. One challenge is shared by all platforms: it can be hard to distinguish real biological changes, such as a small expansion of certain cells during an immune response, from the technical errors that arise from use of the enzymes. E.L.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



GOPAL MURTI/SPL

Adaptive immune cells, such as this mature B cell, can show a staggering diversity of antigen receptors.

So Abate looked at trying a completely different kind of molecular tag: short DNA sequences. The four-letter genetic alphabet can be used to create an enormous number of unique sequences, known as barcodes, for cells or molecules of interest. Abate stained cells with DNA-labelled antibodies and shuttled them into a machine that broke the cells open and spliced the antibody's barcode to a second barcode that identifies the cell from which it came. He could then analyse those linked barcodes using a DNA sequencer. To get single-cell resolution, he turned to microfluidics. His team developed a method called Abseq, which injects individual antibody-stained cells into 10-micrometre droplets that contain unique DNA barcodes for both the antibody and cell[6]. Abseq works more slowly than flow cytometry, but it can survey a theoretically limitless number of proteins in individual cells. And that could help to pinpoint which cells are driving immune activity in various diseases. This is especially important for diseases such as cancer, in which each tumour has unique features that could engage — or suppress — different immune cells.

However, the real power of Abseq lies in its potential to integrate with other single-cell methods, says Sam Kim, a postdoc in Abate's lab who helped develop Abseq. Because it generates information that is readable by sequencing, it can assign multiple types of 'omics data — proteomes, transcriptomes, even disease-causing genetic mutations — to each cell, Kim says.

Mission Bio, a biotechnology company in South San Francisco, licensed the Abseq technology to build just such a multi-omics system. At a meeting of the American Society of Human Genetics in Florida last October, the company launched a first-generation system for analysing gene mutations, and it is now developing protein- and RNA-profiling components. BD Biosciences in San Jose, California, is developing similar technology: last September, it released its Rhapsody system for single-cell profiling of predefined or custom gene panels. Rather than probe the whole transcriptome, BD's technology focuses on several hundred transcripts, which cuts the number of sequencing reads about tenfold, says Nikhil Rao, global product manager for the company's Rhapsody system. "You save a tonne of money if you don't waste sequencing on genes you're not interested in." The company is scheduled to expand the Rhapsody platform to include protein detection with its own AbSeq assays — developed independently of Abate — later this year.

DO-IT-YOURSELF APPROACHES

Single-cell transcriptome analyses can also be used to profile the immune system. These methods offer "a tremendous opportunity to look genome-wide at what a cell is trying to express — what it's thinking at a given moment", says Shalek.

Both commercial and homegrown options are available. In 2015, two teams at Harvard Medical School reported on their development of methods, called inDrop[7] and Drop-seq[8], which use microfluidics devices to analyse RNA expression genome-wide from thousands of single-cell droplets in parallel. A paper published in February[9] describes how to build a compact version of Drop-seq from 3D-printed parts.

On the commercial side, the California company 10X Genomics offers a high-end instrument that can convert cell samples into sequencing-ready data in a day, and Blacktrace Holdings in Royston, UK, offers both an all-in-one system and modular components for building your own.

Still, droplet-based methods have drawbacks. They don't work for clinical samples that contain only a small number of cells. And reagents are wasted when loading cells at low densities to ensure only one goes into each droplet.

Shalek and MIT chemical engineer Christopher Love worked through these issues to create a single-cell RNA-sequencing system called Seq-Well[10], which floats cells onto a silicone array containing 86,000 sub-nanolitre wells. "Think of it as a big ice-cube tray," Shalek says. BD's Rhapsody system uses a similar approach, capturing cells in tiny wells.

A key advantage of Seq-Well is its portability, Shalek says. Members of his lab have taken the device to Thailand to help MIT colleagues analyse cells infected by the malaria parasite *Plasmodium vivax*. They have also trained a team at the Africa Health Research Institute in Durban, South Africa, to use Seq-Well to identify infected immune cells in lymph-node samples taken from people with HIV.

Strategies for collecting multiple 'omics data sets are moving forwards, too. Last year, for instance, two teams independently published methods — called REAP-seq[11] and CITE-seq[12] — for analysing proteins and messenger RNA simultaneously in single cells. The New York University Langone Medical Center and the West German Genome Center are gearing up to offer similar services. Researchers can send us samples, "and we will do the single-cell analyses", says Pratip Chattopadhyay, who directs the New York unit. The German centre, a new multi-institution collaboration headquartered in Bonn, is mobilizing a panel of experts for consultation on the technologies.

"It's a very fast-evolving field with a lot of innovation," says Joachim Schultze, a tumour immunologist at the German Center for Neurodegenerative Diseases in Bonn. "That also means nobody has any idea where this will be in ten years — which technologies will make it, or how many we will have." ■

Esther Landhuis *is a freelance science journalist in the San Francisco Bay Area, California.*

1. Bandura, D. R. *et al. Anal. Chem.* **81**, 6813–6822 (2009).
2. Bendall, S. C. *et al. Science* **332**, 687–696 (2011).
3. Korin, B., Dubovik, T. & Rolls, A. *Nature Protoc.* **13**, 377–391 (2018).
4. Korin, B. *et al. Nature Neurosci.* **20**, 1300–1309 (2017).
5. Cheung, P. *et al. Cell* https://doi.org/10.1016/j.cell.2018.03.079 (2018).
6. Shahi, P., Kim, S. C., Haliburton, J. R., Gartner, Z. J. & Abate, A. R. *Sci. Rep.* **7**, 44447 (2017).
7. Klein, A. M. *et al. Cell* **161**, 1187–1201 (2015).
8. Macosko, E. Z. *et al. Cell* **161**, 1202–1214 (2015).
9. Stephenson, W. *et al. Nature Commun.* **9**, 791 (2018).
10. Gierahn, T. M. *et al. Nature Meth.* **14**, 395–398 (2017).
11. Peterson, V. M. *et al. Nature Biotechnol.* **35**, 936–939 (2017).
12. Stoeckius, M. *et al. Nature Meth.* **14**, 865–868 (2017).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

NEUROSCIENCE A materials scientist tells of his brain-stimulation breakthrough **p.601**

MENTAL HEALTH Workplace support for struggling researchers **go.nature.com/wellbeing**

PODCAST Careers and chronic-illness advice **go.nature.com/chronic**

SHUTTERSTOCK



Researchers who choose to study or work abroad must adapt to customs, hierarchies and expectations that can differ greatly from those they are used to.

WORKING ABROAD

Globetrot for science

How to navigate cultural variances in a lab outside your native country.

BY ROBERTA KWOK

Punam Amratia grew up in Kenya, where she attended a British school. She then completed her bachelor's degree in statistics and a master's in epidemiology in the United Kingdom. In both nations, she was accustomed to people being blunt. If teachers didn't like her work, they were upfront about it. "They told me, 'It's crap. Go back and do it again,'" she says.

So when Amratia started her PhD in malaria epidemiology in 2014, she did not sugarcoat her own opinions. During her first month at the University of Florida in Gainesville, her laboratory met to discuss a paper that her supervisor was reviewing. Amratia recalls calling the paper "shit" and saying that it should not be published. A US lab mate described the paper's positive attributes and outlined items that could be improved. Amratia's colleagues suggested to her that she could be more diplomatic. After the meeting, she tried to temper her directness.

But Amratia also faced problems socializing with her PhD-committee members, partly because she didn't understand US popular-culture references. At a social event, while her colleagues laughed about a popular 1980s TV show, she stayed quiet because she was unfamiliar with it. Such incidents made it hard for her to establish a casual rapport with these faculty members, so she did not feel comfortable asking them for advice during her doctoral programme.

Plenty of opportunities exist to study and work abroad. But some early-career scientists might face challenges adapting to different communication styles and different workplace and academic hierarchies. Supervisors and junior researchers can reduce the risk of misunderstandings by actively learning about each other's cultures and communicating workplace expectations clearly.

It is important both to be sensitive to cultural differences and to avoid inadvertently stereotyping. Nanda Dimitrov, director of Western University's Teaching Support Centre in London, Canada, who has written about cross-cultural graduate supervision, says that it's important not to make assumptions about students solely on the basis of their culture. A wealthy Chinese student from Hong Kong, for example, could see things differently from one who comes from a rural area on the mainland, she notes. And individual perceptions ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

can vary: Amratia says that she personally encountered directness more often in the United Kingdom than in the United States, but others might not have experienced this. Dimitrov points out that the relationship between junior researcher and supervisor is influenced by many factors including personality, previous experiences and the department's and discipline's workplace culture.

Moving abroad to study or work has become more common, partly owing to encouragement from national governments and funding agencies. The number of students pursuing higher education abroad rose from 1.7 million in 1995 to 4.1 million in 2013, according to the *UNESCO Science Report: Towards 2030* (see go.nature.com/2wfvwyq), published in 2015 by the United Nations Educational, Scientific and Cultural Organization (UNESCO) in Paris. European funders promote the practice through programmes such as the Marie Skłodowska-Curie actions, which provide grants for researchers studying or working outside their home countries. From 2003 to 2010, the Chinese government increased the number of scholarships for studies outside China from fewer than 3,000 to more than 13,000. According to the UNESCO report, the regions with the highest proportions of students seeking higher education elsewhere are central Asia, Arab states, sub-Saharan Africa and Western Europe. The most popular destination for all outbound PhD students is the United States, where about half of international science and engineering PhD students are enrolled, followed by the United Kingdom, France and Australia.

Many labs experience few problems with cultural differences. "Science workplaces are so international," says Kaisa Kajala, a Finnish plant biologist at Utrecht University in the Netherlands who has studied or worked in the United Kingdom, Australia and the United States. "People are pretty understanding of different cultures."

When misreadings do arise, however, it's important to address them, because the stakes are often high for international students. If research takes longer than expected, they might have difficulty extending visas or paying tuition fees. Principal investigators (PIs) who make invalid assumptions about a student's intentions might write tepid letters of recommendation for job applications or decide not to collaborate with that person after graduation.

Expectations for behaviour in areas such as leadership, communication and feedback style can vary across cultures. "While it is important never to stereotype, if we don't try and understand how cultures differ, we are missing a huge part of what shapes how people act and how misunderstandings can arise," says Andrew Spencer, director of Rose Window Consulting near Oxford, UK, who has run training programmes for international businesses (including Springer Nature) on managing and communicating across cultural boundaries. For instance, he says, in the Netherlands, negative feedback tends to be given clearly and unambiguously, whereas in Japan, criticism might be more indirect. Spencer recommends *The Culture Map: Breaking Through the Invisible Boundaries of Global Business* by Erin Meyer (PublicAffairs, 2014) for more examples.

One point of difference that can arise is the appropriate level of deference to supervisors. Dimitrov says that some Nigerian, Egyptian and Chinese international students report that in their home nations, a large power differential between students and teachers is common, and that students generally follow instructions without arguing. But a supervisor from a country where debate is expected might sometimes incorrectly interpret a lack of questioning from the student as a lack of interest in the work, says Theresa Winchester-Seeto, an independent consultant on higher education in Sydney, Australia.

"If we don't try and understand how cultures differ, we are missing a huge part of what shapes how people act."

Keshun Zhang faced this issue after moving from China to the University of Konstanz in Germany to pursue his PhD in psychology. He was used to following teachers' suggestions. But "the culture in Germany always encourages you to argue, to fight for yourself", says Zhang, now a postdoc at the university and co-author of the 2016 book *When a Chinese PhD Student Meets a German Supervisor: Tips for PhD Beginners* (University of Konstanz). With his supervisor's encouragement, he started pushing back — for instance, if he thought data should be analysed using a different statistical method, he would say so. After his first year, his supervisor said, "'Wow, in this one year, finally you have learnt to say no,'" Zhang recalls.

Zhang also realized that he was expected to work more independently than he had during his master's programme in China. In Germany, he peppered a postdoc with frequent questions about issues such as statistical methods. His supervisor urged him to try to solve problems on his own and to seek guidance only if he became stuck. Zhang initially found this approach difficult but came to prefer it.

The absence of a strictly defined hierarchy can encourage freer communication, says Salim Reza, a radiation-detector scientist at Mid Sweden University in Sundsvall. When he moved from his native Bangladesh to Sweden for graduate studies, he learnt that he did not need to address faculty members as 'sir' or 'professor', or remain standing in their offices. This informality made it easier for him to approach professors to clarify a topic or to propose a new research angle. "I could go with the craziest idea to my teacher, and he would explain why it was good or bad," he says.

Supervisors also can clarify expectations for the student's responsibilities when they arrive. The *Western Guide to Graduate Supervision* (University of Western Ontario Teaching Support Centre, 2008) includes a rating scale that asks questions such as whether the student or supervisor will select the research topic and decide on methods. A student and PI could fill out the form together, discuss their answers and resolve discrepancies, Dimitrov says.

Davor Solter, a mammalian-developmental biologist from Croatia who now splits his time between New Mexico and Maine, says that he never encountered problems owing to cultural differences as a PI in the United States, Germany and Singapore. His lab members' countries of origin were irrelevant to their performance, he says.

MIXED MESSAGES

Sometimes, though, misunderstandings can stem from differences in communication style. In some countries, the 'feedback sandwich' is common, Dimitrov says: start with praise, suggest improvements and end with encouragement. Students from countries where this format is less common might think that because comments were mostly positive, the suggestions are optional and can be ignored, she says. To avoid such mishaps, students could write an e-mail after each meeting summarizing the feedback and next steps so that the supervisor can correct their interpretation if needed.

Conversely, a student who is accustomed to gentler feedback might be 'traumatized' by cultural tendencies in other countries, such as Germany or the Netherlands, to give more direct criticism, Dimitrov says. Senior researchers could smooth over differences by discussing how the student prefers to receive comments, she suggests. Students could also talk to lab mates about the feedback; hearing others' stories of critiques — and positive outcomes, such as getting a paper accepted — could help them to overcome discouragement.

Some researchers are surprised by their host country's e-mail practices. Laetitia Wilkins, a postdoc in evolutionary biology



HERO IMAGES/GETTY

Communication styles can vary across cultures.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

at the University of California, Berkeley, says that in her native Switzerland, people typically reply to nearly every work-related message. "It's almost mandatory that you answer," she says. In California, she found a more relaxed approach. When she e-mailed another researcher to request a data set and heard nothing, she wondered if she had done something wrong and the person did not want to collaborate. She did not send a follow-up message because, she says, that would be considered rude in Switzerland. Eventually she realized that ignoring or 'losing' e-mails was common in the United States. Now, if a message goes unanswered, she sends another e-mail a few days later and usually gets a friendly response.

Other scientists find that the priorities attached to socializing differ from what they are used to. In Reza's experience, people in Bangladesh do not usually take coffee breaks at work. But in Sweden, he realized that it was important to attend *fika*, coffee breaks that the department usually took each day. "We are obsessed with *fika*," he says. The gatherings provide opportunities to make social plans, hear what other groups are doing and discuss research issues.

Some scientists welcome the change of pace offered by a fresh environment. Ecologist Christine Lucas moved from the United States to Uruguay for a postdoc position and is now a faculty member at the University of the Republic in Paysandú. She found that her new colleagues tended to keep a better work–life balance. When she became pregnant, she easily postponed the start of her postdoc by a few months; once she began, she had flexible hours, could sometimes work from home and designed a project that required little fieldwork. Friends who did postdocs in the United States seemed, to her, under more pressure to work with lab members face-to-face and to meet tough publication targets. And, she adds, it is acceptable for her department meetings in Uruguay to start 5–10 minutes late — allowing for unexpected obstacles such as transport strikes.

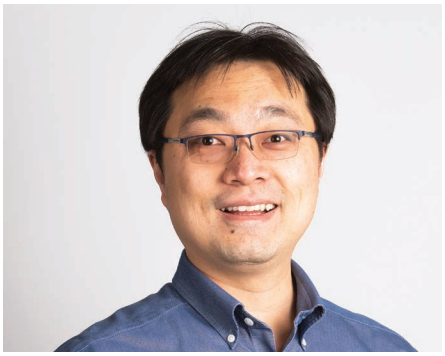
Whether they are welcoming international students or starting work in new countries, scientists can ease the transition by remaining non-judgemental. People sometimes brush off a student from another country as 'rude', but "in their culture, they're not", Amratia says. Researchers should also remember that their nation's customs aren't necessarily best. Solter says that his Croatian background helped him here: "When you come from a small country, you don't assume everybody should be doing things your way," he says. "I never cared if somebody was different than me as long as it didn't seriously affect the rest of the lab." ■



Roberta Kwok *is a freelance writer in Kirkland, Washington.*

BACK STORY
Brain stimulator

Materials scientist Bozhi Tian developed silicon-nanowire solar cells in 2007. Now, at the University of Chicago in Illinois, he has shown how similar nanoscale devices can be used to manipulate brain signals with light.



JEAN LACHAT/UNIVERSITY OF CHICAGO

Describe what your team's latest study has achieved.
We've shown that we can stimulate behaviour in a mouse by placing a nanoscale silicon mesh on the part of the brain that translates nerve impulses into movement, and then shining a light onto the mesh (Y. Jiang *et al. Nature Biomed. Eng.* https://doi.org/10.1038/s41551-018-0230-1; 2018). For example, we were able to control the animal's left forelimb by flashing a beam at the right side of its brain.

Why is this approach game-changing?
Until now, there have been only two main methods of neurostimulation. The first involves implanting electrodes to produce neural activity. The second, called optogenetics, requires cells to be genetically modified so that they can be controlled using light, but genetic modification is difficult and carries ethical concerns. Our approach uses a non-genetic neurostimulation device that enables distant nerves to be stimulated with light. It could have a big impact on the treatment of pain and other disorders.

Has it led your research in a new direction?
Yes, we now plan to do research on non-human primates — for example, exploring how to restore grasping functions after paralysis.

Does the device have potential clinical applications?
If we put the silicon mesh on the brain surface, we can elicit some activity deep inside the brain. And if you can stimulate the brain, you can treat certain neurological diseases or help a person to regain control over parts of their body. So it could be used to treat diseases such as Parkinson's or disorders such as depression.

Describe one of your breakthrough moments.
My graduate student Ramya Parameswaran has demonstrated how a silicon nanowire can stimulate single cells (R. Parameswaran *et al. Nature Nanotech* **13**, 260–266; 2018). Some of the gold that catalyses growth of the silicon wire becomes individual atoms that cover the wire's surface. When we removed the atomic gold, some neurons couldn't be activated. We later found that the gold enhanced the silicon's electrochemical properties and made it a better neurostimulator.

You switched your research focus from energy to neurostimulation. What happened?
I left China to start a PhD at Harvard University in Cambridge, Massachusetts, in 2004, working on the use of single nanowires in photovoltaics. We showed that silicon nanowires can convert sunlight into electricity, just as conventional solar cells do. However, they are also small enough to be integrated into nanodevices. After that work was published (B. Tian *et al. Nature* **449**, 885–889; 2007), I decided to completely change my research interest from energy to electrophysiology and bioscience.

Why did you make that change?
Biology offered lots of opportunity and room for exploration, especially in bringing nanowires and neuroscience together. And I like to explore the unknown. So, for the second half of my PhD, I worked on developing a transistor that's small enough to be delivered into a single cell to record electrical activities.

How was the transition?
It was tough. I had no experience in cell culture, electrophysiology or working with animal tissues. I took a neuroscience course and learnt tremendously from that. Even so, there were setbacks. For example, my cell cultures kept getting contaminated. But I told myself that good moments would come if I persisted.

How did you find collaborators?
When I began at Chicago, some biologists were interested in collaborating; others weren't. I gave a lecture to biophysics students, and Ana Correa, a biochemist and the wife of molecular biologist Francisco Bezanilla, attended. She introduced me to her husband afterwards, and he and I have since written grants and papers together. Sometimes you just need that right moment and right person. ■

INTERVIEW BY VIRGINIA GEWIN

This interview has been edited for clarity and length.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

DNA EXCHANGE

The course of true love.

BY D. A. XIAOLIN SPIRES



ILLUSTRATION BY JACEY

I placed the dishes in a haphazard pile in the sink and watched as the auto-sorter pulled them into the dishwasher one by one, stacking them into supersonic racks. My mind reeled from the phone call from school. "Meiling was acting up again." I mouthed *start* and the DishU hummed, giving off a scent of lemon in the process. The soothing movements of Smart DishU filled my eyes, mechanical limbs like spiders cartwheeling across my porcelain, singing them clean.

I was lost in thought when I heard Securo's deep voice announce: "Meiling has arrived." The door slammed closed. I heard heavy footsteps run up the stairs.

"Meiling, get over here," I yelled. The footsteps paused, then I heard them in reverse. Trudge, trudge. Down and across the living room.

She had her baseball cap on, which flashed 'Just 4 UR Evite', the newest band sensation. Who were they again? Josh, Kick and Enlai? Or was that last week's boy band?

"Take that thing off and look at me," I said. She shook her head, leaving a static holotrail ghost of the projection, her eyes hidden under the rim.

I pulled off her cap. She shrieked. Her black hair flew in all directions. *Teenage years, when would they be over?*

As she thrashed, saying it was *unfair*, I saw something jiggling off her neck, swaying with her angry convulsions.

"What is *that*?" I pointed. It was flesh-coloured and squishy, a nub. It looked like a dead naked mole rat, hunched in fetal position.

Meiling's eyes lit up, looking utterly delighted. Her tantrum stopped.

Dangling off her neck on a chain was a deformed, pale locket.

She picked up the piece, squishing it slightly. The pendant compressed under the pressure of her finger pads.

"You know you say I should stop being so materialistic," Meiling said, smirking. "Well, I told Kyle that. He wanted to go steady with me and told me he'd buy me anything. I said, 'I don't need you to buy me anything, but if you're so into me, why don't you give me your right ear?'"

She laughed. It was cruel, the sound of my own narcissistic self echoing in my firstborn.

"He said, 'Ear? You mean, you want me to hear for you? I could install a chip and translate for you in Spanish class.' And I said, 'No, I don't need you to do my schoolwork, thanks. I want your ear, literally.' He smiled, thinking it was a joke."

"His ear," I said, incredulity cracking out of my voice.

"Yeah, mom, keep up," she said, rolling her eyes. Her fingers worked her way into the seams and folds of the whorl of pale dough in her hands.

"Well, he did it. He gave me his ear. I thought it was so romantic, like Van Gogh from art class."

"That's his ear, right there?" I felt queasy, my lunch was stirring in my stomach. I saw it, the whorls, the lobe. It felt like the air sucked up all sound. Only the DishU going off in a hum.

"No, mom, don't be dumb."

"Don't you call your mother that." I reached for her pendant, but pulled back at the same time, unsure whether I wanted to touch the supple cartilage.

"He *grew it*. Really, it had another purpose, for the science fair, but he grew an extra one for me. On his arm, he'd been wearing long sleeves even in this June heat, because he wanted to surprise me. They sprouted on his forearm, right above the wrists."

She pointed at her own wrists. I scoured them, checking for any weird nubs.

"That doesn't make it much better."

"He said it doesn't have the organs to actually hear yet, but he's working on the eardrum and seeing if he can project aural interpretation into his room. That way we can be connected."

I know teens have their own courtship rituals, but this was out of control.

"Give me that." I pulled at her chain.

"No!" she screamed. She seized the chain. "This is his promise necklace."

We struggled for five minutes.

She shouted, "You could never understand me" and "Why did I even bother telling you" and "I should've just lied."

I yelled, "You just like to act out to get my attention" and "Well, now you have it and you better get rid of that gross piece of human meat off your chest." The fresh scent of lemon continued to fill the room, a strength escalating in tune with our voices.

"Mom, you'll never understand. You never lived with DNA exchanges and flitskin-tats, you'll never know how much this means!" She threw herself around to rush out.

In her haste, her shirt behind her flew up and for a second, I saw an orb-like lump.

"What is that?" I asked, my voice now low, unusually calm.

Meiling knew that tone. That was the tipping point. If she pushed me over, she'd be grounded for weeks. No holocalls, no late-night trips to dessert shack, no e-cart creds.

"It's, uh, a decorative thing. Temporary." She pulled her shirt down.

I walked over and took a look. Her lips were quivering.

It was an eye. A single eye, protruding right above her hips, resting on her lower back like a tramp stamp.

"No, no, no…"

"Well, you know. I can't leave him hanging. Love goes both ways. And he always said what beautiful eyes I had," she said, squirming.

I looked into the pupil, which followed my face. It blinked. Its sinewy veins looked like DishU mechanical legs, spidery threads spreading across her back. I grabbed the holophone and mouthed "Dermatologist", wondering how much this removal surgery would set us back. ■

D. A. Xiaolin Spires *spins in tune to the cosmos. Work in* Clarkesworld, Analog, Fireside, Grievous Angel, Galaxy's Edge, LONTAR, Terraform, Sharp & Sugar Tooth *and* Broad Knowledge. *Web: daxiaolinspires.wordpress.com Twitter: @spireswriter.*

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



ILLUSTRATION BY NICK REDDYHOFF

Building the future

Hong Kong is using its proximity to science hubs in mainland China to encourage innovation, but barriers to collaboration remain.

BY JACK LEEMING

In March this year, Hong Kong officials earmarked 50 billion Hong Kong dollars (US$6.37 billion) for investment in scientific and technological innovation. The move signalled a desire for tighter connections with China's new science-based economy, and wish to diversify Hong Kong's own.

Stephen Phillips runs a government department called InvestHK that promotes investments into the territory from the rest of the world. He hailed the budget announcement as "tangible evidence of the government's determination to position Hong Kong well in a very competitive landscape".

This is a divergence from the history of the region. Since 1997, when the UK government returned the territory to China after its 99-year lease expired, Hong Kong has relied mainly on the historic strengths of its banking and property sectors to drive its economy (it has the world's least-affordable housing market). This has left it vulnerable to shifting economic tides such as the 1997 Asian financial crisis. The rest of China, meanwhile, pivoted towards science and technology to bolster its economy, which was previously known for low-cost manufacturing on a vast scale.

Shenzhen, for example, a city just to the north of Hong Kong, which itself sits on a narrow peninsula in the Guangdong province of southern China, currently commits more than 4% of its gross domestic product (GDP) to research and development (R&D), much of that from private companies. Shenzhen is home to vast science firms, including BGI, the world's largest genomics company, and DJI, the world's market leader in drone technology. Hong Kong dedicated only 0.79% of its GDP to R&D in 2016.

Increasingly, collaboration with the mainland, especially Shenzhen and the surrounding metropolitan area, is the key to maintaining ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Hong Kong's economic strength in the future, says Matthew Evans, who runs the faculty of science at Hong Kong University.

The Pearl River Delta, of which the territory is a part, is home to 66 million people. The region is named after the eponymous waterway that splits the area in half and forms Hong Kong island. "If you lift Hong Kong out and put it in the middle of the Pacific Ocean, will it still perform? Obviously not," explains Andrew Leung, a business consultant on Chinese issues, who is based in the city. "Hong Kong must link to somewhere."

Before the increased budget, Hong Kong had already had some success in science through its existing infrastructure. Last month, for instance, the company SenseTime was crowned the world's most valuable artificial-intelligence (AI) start-up, following an investment from Alibaba, the e-commerce giant, of more than US$600 million.

Originally conceived at the Chinese University of Hong Kong (CUHK), SenseTime transferred into the territory's Science Park system, a government-financed incubator that offers subsidized benefits to start-ups.

SenseTime's speciality is using AI technology for facial recognition, and is regularly criticized by privacy campaigners. In the mega-metropolis of Guangzhou, also part of the Pearl River Delta, this method is used by local police to track the faces of suspected criminals. The company says it has helped to solve at least 100 cases in the city.

As well as the budget announcement, Leung highlights other movements to spur scientific research in the territory in recent years, including the opening of the Innovation and Technology Bureau. It was established in 2015 and is headed by Nicholas Yang, who was formerly vice-president of Hong Kong Polytechnic University. "The question," says Leung, "is how to marry research with businesses."

WAYS TO COLLABORATE

Of the 50 billion Hong Kong dollars from the city's budget, 20 billion will be spent developing a physical manifestation of the spirit of collaboration. The Lok Ma Chau loop — an area of farmland between Shenzhen and Hong Kong — has been tagged by both cities' officials as the site of a science and technology park. Officials say the park will be four times the size of current facilities, and it will be managed by the Innovation and Technology Bureau. Feedback is varied: critics have labelled it a government scheme to buy and develop property rather than anything truly committed to promoting R&D.

Others disagree. Representatives for the science park declined to be interviewed for this article, but from 2016 to 2017 the number of start-ups enrolled in its incubator programme, which offers technical and professional support and subsidized rents to science businesses, increased by 8.4% from the one-year period before.

Biochemist Vasu Saini, who moved to the city from India in 2016 to start his PhD at Hong Kong University of Science and Technology (HKUST), says he can see the value of the science park when he considers the jobs available to his peers. "Hong Kong is driven mostly by the economic sector so there are not a lot of jobs for scientists," he says. "But now there's a science park, there are more opportunities for people with that background."

COLLABORATIVE SPIRIT

In 2010, BGI established a facility in Hong Kong to supplement its existing sites in Shenzhen and form the basis of a collaboration with CUHK. This was a welcome addition to the employment landscape for some. "I didn't really know if I could find a science job in Hong Kong at first, but there are a lot of new companies and there's a lot of new science here," says Irene Chik, who was born in Hong Kong, and manages a lab in the facility. Her family emigrated to Canada in 1996 and she returned to Hong Kong in 2013.

POLITICS IS ANOTHER BARRIER TO FURTHER COLLABORATION AND INTEGRATION.

Academic work is also crossing territorial and cultural borders. Although Hong Kong has five universities in the top 100 QS rankings for 2018, "there's not normally much of a market for the basic research that Hong Kong universities conduct", says Naubahar Sharif, who researches innovation and technology policy at HKUST. He says that collaboration with companies elsewhere in China might help to transfer that basic research into commercial applications, and he adds that collaboration with Shenzhen "allows the university sector in Hong Kong to have an outlet".

One example of this comes from DJI, which was founded in 2006 and, like SenseTime, was conceived in a Hong Kong university and expanded into Shenzhen when the time came to scale up production. Kei May Lau, an electronics engineer at HKUST, worked in the lab next door to robotics engineer Zexiang Li, who was an early adviser and investor at DJI. One day, Li showed her some of the images he'd taken with a drone he'd made with Frank Wang — now the billionaire chief executive of DJI. The company's drones have become ubiquitous worldwide, especially in film-making and photography. "There were these scenic mountains that they climbed up" and flew the drone over, says Lau. "I said, 'Wow'. In the old days, if *National Geographic* wanted to do this, they'd have to spend thousands of dollars on a helicopter to get there. Now, they can take everything they need in a backpack."

Other universities are sitting up and taking notice of these and other commercial successes, and a rush to work with Chinese collaborators is starting in earnest, partially facilitated by incentives from the government in mainland China to encourage collaboration through the Hong Kong–Shenzhen border and across academic–industry barriers. These include giving researchers access to federal funds previously unavailable to scientists in Hong Kong, a development publicized in Chinese state media this month. An older scheme to share the costs and results of applied research with private-sector partners is another positive.

Lau says that this has resulted in funding in Hong Kong being skewed towards applied research. One of her projects is funded by the Innovation and Technology Bureau, which insists that scientists must match their funding with the same amount of investment from the private sector. "They want to make sure you're not doing pie-in-the-sky research," she says.

Companies in China and universities in Hong Kong can both find natural benefits in collaboration, says Sharif. "Shenzhen and the nine or so other cities in the Pearl River Delta have been called the workshop of the world, and they have expertise in manufacturing which Hong Kong doesn't — but Hong Kong has these world-class universities which the delta doesn't have."

Evans sees the advantages for institutions such as his own, but says that any significant collaboration will take an investment of time and good faith.

"One of the challenges for this university and for the rest of Hong Kong is to make these collaborations work," he says. "We're part of the Pearl River Delta too, and we should be one of the driving forces on the academic side of that. If I have anything to do with it, we will be."

"That's only going to work if the academic sector engages with the business sector," he says. "I'm spending increasing amounts of my time going over into that area, into mainland China, and talking to companies and people there about how we can work together."

FEELING FAMILIAR

Phillips points to Hong Kong's Western-style legal framework and intellectual-property laws — hangovers from its history as a British colony that were enshrined in the 'one country, two systems' agreement drafted in the reunification process with China. Overseas investors are comfortable spending in Hong Kong because it is culturally accessible, he argues, unlike the Chinese mainland, where many feel that the financial risk is too great and the landscape too unfamiliar, and where

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



The China–Hong Kong Nature Science Project in December 2015.

JEROME FAVRE/EPA/REX/SHUTTERSTOCK

foreign companies often find themselves mired in bureaucracy. That works both ways. Hong Kong is also a place for Chinese businesses — in science and elsewhere — to internationalize. Vivek Wadhwa, a distinguished fellow at Carnegie Mellon University's campus in Mountain View, California, and a columnist on scientific innovation, agrees. "The West doesn't trust Chinese companies. But it does trust Hong Kong's laws," he says. "This solves many problems for the West."

Culturally, too, Hong Kong has draws for Western investment and talent that mainland China lacks. "If I take you and drop you down in most Chinese cities, the road signs to the food to the language to the culture will be overwhelmingly Chinese," says Evans. "Hong Kong is a more accessible city for westerners."

David Zweig studies Chinese diplomacy and the country's efforts to attract researchers at the Center on China's Transnational Relations at HKUST, which he runs. He says that there are more differences between China and Hong Kong when it comes to their academic sectors. "The level of academic freedom from bureaucratic interference is preferable in Hong Kong. You don't need to kiss ass to government administrators so much," he says. He says that returnees, people born in China who studied and worked elsewhere and have since returned, complain that the bureaucrats "have too much power in the mainland; if you want to get access to research money in your first couple of years, it's not easy."

There are still barriers to collaboration, however. Wadhwa says that although Hong Kong might have aligned itself with China's science and technology priorities, this is not necessarily enough to maintain economic growth. "Yes, there might be opportunities to collaborate with places like BGI in Shenzhen," he says, "but so what? That doesn't build a new economy for Hong Kong. That's not true innovation."

Politics is another barrier to further collaboration and integration. Many of Hong Kong's citizens remain protective of their rights to vote, publish and march against the government's interests — rights that those outside the city do not have. The success of data companies such as SenseTime, which critics say will cement China as a surveillance state, does little to alleviate common fears about what life will be like as we approach 2047 — the due date for full integration with China.

PROGRESS

Whatever happens politically, it's likely that scientists in Hong Kong and the rest of the Pearl River Delta will draw closer together as time goes on — and many in the scientific world generally see this as a positive. "When you look at what's happening in this neck of the woods, it's about complementarity," says Phillips. "It's not about one city being set up against another. That's not how businesses see it. They look at how they can deliver business."

Evans is confident that collaborations across the border will strengthen in academia, too. "We have to work hard to make it happen now," he says. "If it doesn't happen now, we're going to lose our place." Fortunately, Hong Kong is a city used to swift transformation, Chik says, especially when compared with her previous home of Vancouver. "It's fast-paced and ever-changing, and if you don't change fast enough, then all these competitors — they will overtake you." ■

Jack Leeming *is the editor of Naturejobs.*

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.